РОССИЙСКАЯ АКАДЕМИЯ НАУК
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР ИМ. А. А. ДОРОДНИЦЫНА РАН

На правах рукописи

КОЧЕДЫКОВ ДЕНИС АЛЕКСЕЕВИЧ

ОЦЕНКИ ОБОБЩАЮЩЕЙ СПОСОБНОСТИ НА ОСНОВЕ ХАРАКТЕРИСТИК РАССЛОЕНИЯ И СВЯЗНОСТИ СЕМЕЙСТВ ФУНКЦИЙ

05.13.17 — теоретические основы информатики

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель
д.ф.-м.н. К. В. Воронцов

Москва, 2011

Оглавление

Глава 1

Введение

Работа посвящена проблеме повышения точности оценок обобщающей способности в задачах обучения по прецедентам.

Актуальность темы. В задаче обучения по прецедентам рассматривается *генеральная совокупность* объектов на которой задана некоторая *целевая* функция. Из совокупности случайным образом извлекается *обучающая выборка* объектов (прецедентов). *Метод обучения* получает на вход обучающую выборку со значениями целевой функции на ее объектах и на выходе дает функцию, которая должна аппроксимировать целевую функцию на оставшейся (скрытой) части генеральной совокупности, называемой *контрольной выборкой*. Функцию, возвращаемую методом, традиционно называют *алгоритмом*, имея в виду что процедура вычисления значения функции на объектах совокупности должна допускать эффективную компьютерную реализацию. Множество всех алгоритмов, которые может выдать метод обучения, называют *семейством алгоритмов*.

Задача оценивания обобщающей способности метода обучения состоит в том, чтобы, имея в распоряжении лишь наблюдаемую обучающую выборку, определить, насколько хорошо алгоритм, выданный методом, аппроксимирует целевую зависимость на скрытой части совокупности. Качество алгоритма на множестве объектов обычно характеризуется числом или частотой ошибок аппроксимации. Если частота ошибок на скрытой части (частота на контроле) существенно выше, чем на обучающей выборке (частота на обучении), то говорят, что метод переобучился или что выбранный им алгоритм переобучен.

В предположении, что все обучающие выборки заданного размера равновероятны, ставится задача оценивания *вероятности переобучения* метода. В англоязычной литературе такая постановка для бесконечной генеральной совокупности носит на-

звание *PAC-обучения* (Probably Approximately Correct learning, [63, 19]) и является развитием теории Вапника-Червоненкиса [65]. Случай конечной генеральной совокупности рассматривается в *теории надежности обучения по прецедентам* [5].

Чтобы исключить зависимость от метода обучения, имеющего обычно довольно сложную структуру, рассматривают *вероятность равномерного отклонения частот* — вероятность того, что в семействе *возможно* выбрать алгоритм у которого частота ошибок на контрольной выборке существенно больше его частоты ошибок на обучающей выборке.

Классические оценки в PAC-теории завышены, поскольку ориентированы на худший возможный случай целевой зависимости. Одним из наиболее актуальных направлений исследований в связи с этим является получение оценок, зависящих от свойств целевой функции, семейства и обучающей выборки.

Одними из основных факторов завышенности классических оценок является пренебрежение *расслоением* семейства по частоте ошибок и *сходством* алгоритмов в семействе. Учет обоих факторов приводят к существенному уточнению оценок вероятности переобучения в комбинаторной теории [Воронцов, 2009]. Однако точные оценки к настоящему моменту получены лишь для некоторых модельных семейств алгоритмов и довольно узкого класса методов обучения.

Цель работы. Разработка новых методов получения оценок обобщающей способности, учитывающих расслоение и сходство для произвольных семейств и методов обучения, в рамках комбинаторной теории надежности обучения по прецедентам.

Научная новизна. В работе развиваются два метода получения оценок обобщающей способности. Оценки, использующие расслоение семейства по частоте ошибок, рассматривались ранее в контексте классической PAC-теории. В данной работе выводятся их комбинаторные аналоги с некоторыми улучшениями. Второй метод — оценки, учитывающие сходство алгоритмов в смысле расстояния Хэмминга между векторами ошибок алгоритмов. Он основан на неравенствах типа Бонферрони, оценивающих вероятность конъюнкции большого числа событий через вероятности дизъюнкции их различных комбинаций и технику производящих функций. Данный метод является новым. Основная оценка параграфа 4.5 вводит понятие *степени связности* алгоритма a — числа алгоритмов на единичном расстоянии от a и понятие *профиля связности* семейства — распределение степени связности в семействе. Оценка улучшает базовую оценку Вапника-Червоненкиса на множитель, экспоненциальный

по средней степени связности алгоритмов в семействе (для линейных классификаторов — по размерности пространства параметров).

Для семейства линейных классификаторов в работе получены оценки среднего значения и дисперсии профиля связности.

Методы исследования. Основными методами исследования в работе являются методы комбинаторной теории надежности обучения по прецедентам, оценки концентрации вероятностной меры, неравенства типа Бонферрони-Галамбоса, используемые для оценивания вероятности равномерного отклонения частот, метод производящих функций перечислительной комбинаторики, используемый для вычисления отдельных слагаемых неравенств типа Бонферрони. Для анализа профиля связности семейства линейных классификаторов в работе используется теория геометрических конфигураций, применяемая в теории обучения для решения гораздо более простой задачи — оценивания числа алгоритмов в семействе. Для экспериментального вычисления и сравнения оценок используется метод Монте-Карло.

Теоретическая и практическая значимость. Работа носит в основном теоретический характер и вносит существенный вклад в развитие комбинаторной теории надежности обучения по прецедентам. Предложенные методы учета сходства алгоритмов могут применяться для конкретных семейств как в рамках комбинаторного подхода, так и в рамках классического PAC-подхода для уточнения оценок, использующих неравенство Буля.

Основным практическим приложением оценок обобщающей способности является разработка и обоснование новых методов обучения. Они также могут служить источником качественных соображений при выборе семейства. К примеру, основная оценка настоящей работы показывает, что при повышении сложности семейства, существенным фактором, уменьшающим вероятность переобучения, является увеличение степени сходства алгоритмов в семействе, что может служить обоснованием для применения семейств, непрерывных по параметрам.

Область исследования согласно паспорту специальности 05.13.17 — «Теоретические основы информатики»:

- разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных (п. 5);
- моделирование формирования эмпирического знания (п. 7);

- разработка методов обеспечения высоконадежной обработки информации (п. 11).

Согласно формуле специальности «Теоретические основы информатики», к ней относятся, в числе прочего, «. . . исследования методов преобразования информации в данные и знания; создание и исследование. . . методов машинного обучения и обнаружения новых знаний. . . ». Таким образом, исследование проблемы переобучения соответствует данной специальности.

Апробация работы. Результаты работы докладывались на научных семинарах ВЦ РАН, на конференциях ММРО, ИОИ, научной конференции МФТИ:

- всероссийская конференция «Математические методы распознавания образов» ММРО-12, 2005 г. [10];
- международная конференция «Интеллектуализация обработки информации» ИОИ-6, 2006 г. [8];
- всероссийская конференция «Математические методы распознавания образов» ММРО-13, 2007 г. [11];
- научная конференция МФТИ 50 «Современные проблемы фундаментальных и прикладных наук» 2007 г. [9];
- научная конференция МФТИ 51 «Современные проблемы фундаментальных и прикладных наук» 2008 г. [6];
- всероссийская конференция «Математические методы распознавания образов» ММРО-14, 2009 г. [7].

Результаты неоднократно докладывались на семинарах отдела Интеллектуальных систем ВЦ РАН, (рук. чл.-корр. РАН Константин Владимирович Рудаков).

Публикации по теме диссертации в изданиях из списка ВАК: [41, 42]. Другие публикации по теме диссертации: [10, 8, 11, 9, 6, 7]. Текст диссертации доступен на странице автора `www.MachineLearning.ru/wiki`, «Участник:Denis Kochedykov».

Структура и объём работы. Работа состоит из оглавления, введения, пяти глав, заключения, списка обозначений, списка литературы.

Краткое содержание работы по главам. Во введении определяются основные обозначения, определяется функционал вероятности переобучения и процедура его обращения для получения доверительных оценок частоты ошибок на контроле. Определяется квантильный критерий переобучения. Приводятся комбинаторные аналоги оценки Вапника-Червоненкиса (VC-оценка) и оценки «бритвы Оккама» и выводится оптимальное априорное распределение на множестве алгоритмов для оценки «бритвы Оккама».

Глава 2 содержит краткий обзор некоторых известных методов и результатов теории статистического обучения.

В главе 3 выводятся комбинаторные аналоги оценок обобщающей способности учитывающих расслоение семейства – так называемых shell-оценок [48, 47], основанных на следующем соображении. Обычно бóльшая часть алгоритмов в $\mathcal{F}$ имеет вероятность ошибки (или частоту ошибок на полной выборке) около 50%. Если метод обучения выбирает алгоритм с малой частотой ошибок на обучающей выборке, то фактически выбор производится не из всего семейства алгоритмов, а лишь из небольшой его части, состоящей из алгоритмов с малой вероятностью ошибки. Размер этой части семейства существенно ниже размера всего семейства, что предполагает возможность точнее оценить вероятность переобучения по сравнению, скажем, с классическими оценками Вапника-Червоненкиса. В параграфе 3.3 выводится комбинаторная shell-оценка, которая учитывает эффект расслоения семейства по частоте ошибок на полной выборке, но требует знания полной выборки, то есть является ненаблюдаемой. В параграфе 3.4 выводится наблюдаемая комбинаторная shell-оценка, основанная на расслоении семейства по частоте ошибок на обучающей выборке.

В главе 4 выводятся оценки обобщающей способности, учитывающие сходство алгоритмов в семействе. Оценки основаны на следующих соображениях. В VC-оценке функционал равномерного отклонения частот оценивается сверху при помощи неравенства Буля. Известно, что оно может быть сильно завышено при большом числе входящих в него событий. Неравенство, очевидно, становится точным, только если все события в нем взаимоисключающие. При оценивании функционала равномерного уклонения частот, составляющие его события вида «алгоритм a переобучился», наоборот, существенно совместны и их число в неравенстве крайне велико. Как следствие, неравенство Буля представляет собой один из самых существенных факторов завышенности VC-оценки. Можно улучшить оценку неравенства Буля учитывая пересечения входящих в него событий. Пример такого учета дает разложение функционала равномерного отклонения частот по принципу включения-исключения. Очевидно, что пересечение событий вида «алгоритм a переобучен» составляющих функ-

ционал равномерного отклонения частот определяется сходством соответствующих алгоритмов. В главе рассматривается хэммингово сходство векторов ошибок алгоритмов, определяется граф 1-сходства на множестве векторов ошибок, предлагается способ вычисления вероятностей пересечений соответствующих событий в функционале равномерного отклонения частот и выводятся оценки, учитывающие различные характеристики сходства алгоритмов в семействе. В параграфе 4.4 выводится оценка учитывающая связность графа 1-сходства семейства, в параграфе 4.5 выводится оценка учитывающая распределение полустепеней вершин в графе — профиль связности, в параграфе 4.6 выводится оценка учитывающая наличие цепей в графе.

В главе 5 исследуются характеристики профиля связности для семейства линейных классификаторов. Поскольку точность оценок предшествующей главы увеличивается с ростом степени связности семейства, возникает вопрос о нижних оценках степени связности или оценках степени концентрации профиля связности для конкретных семейств. Такие оценки можно было бы использовать в оценках обобщающей способности вместо профиля связности. В главе при помощи теории конфигураций выводится точное среднее значение и оценка дисперсии полустепени связности для семейства линейных классификаторов. Полученные оценки не зависят от генеральной совокупности и представляют комбинаторные характеристики семейства линейных классификаторов.

В главе 6 экспериментально сравниваются оценки обобщающей способности из предшествующих глав. Используется семейство линейных классификаторов и случайная генеральная совокупность. Профили расслоения и связности оцениваются по методу Монте-Карло, для этого предлагается и обосновывается процедура сэмплинга алгоритмов из семейства.

1.1 Обозначения

Пусть $\mathcal{X}$ некоторое множество объектов и Y некоторое множество «ответов» допустимых для объекта из $\mathcal{X}$. Пусть $\mathcal{F}$ некоторое параметрическое семейство отображений из $\mathcal{X}$ в Y, будем называть их *алгоритмами*, имея ввиду, что они должны допускать эффективную вычислительную реализацию. Пусть $\mathbb{X} \subset \mathcal{X}$ некоторая фиксированная конечная генеральная совокупность из L объектов, будем называть ее *полной выборкой*. Пусть задана бинарная *функция потерь* $I\colon \mathcal{F} \times \mathbb{X} \to \{0,1\}$. Если $I(f,x) = 1$, то говорят, что алгоритм $f \in \mathcal{F}$ допускает ошибку на объекте $x \in \mathbb{X}$.

Например, в задачах классификации обычно полагают $I(f,x) = [\,f(x) \neq y(x)]$; в задачах регрессии можно полагать $I(f,x) = \Big[\,|f(x) - y(x)| > \gamma\Big]$. Здесь и далее

квадратные скобки обозначают индикаторную функцию, равную 1, если условие в скобках истинно, и 0 иначе.

Рассмотрим конечное множество L-мерных бинарных *векторов ошибок* алгоритмов из $\mathcal{F}$ на полной выборке $\mathbb{X}$:

$$A = \left\{ \big(I(f, x_i)\big)_{i=1}^{L} \;\middle|\; f \in \mathcal{F} \right\} \;\subseteq\; \{0,1\}^L . \tag{1.1}$$

Допуская несущественную нестрогость в обозначениях, будем далее для краткости обозначать (1.1) как $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$.

Определим отношение эквивалентности на $\mathcal{F}$: $f_1 \sim f_2$, если $I(f_1, x) = I(f_2, x)$ для всех $x \in \mathbb{X}$. Элементы множества $a \in A$ взаимно однозначно соответствуют классам эквивалентности на $\mathcal{F}$. Будем говорить, что $f \in \mathcal{F}$ *представляет* алгоритм $a \in A$, если $a = (\mathrm{I}(f, x_1), \ldots, \mathrm{I}(f, x_L))$.

Для краткости везде далее будем называть векторы ошибок из A также алгоритмами, имея ввиду произвольный алгоритм из соответствующего класса эквивалентности на $\mathcal{F}$.

Заметим, что мощность A может быть оценена сверху либо как $\sum_{k=0}^{d} \binom{L}{d}$, где d есть VC-размерность [64] семейства бинарных функций $\{I(f, \cdot)\colon f \in \mathcal{F}\}$, либо как 2^L, если его VC-размерность бесконечна.

Число ошибок алгоритма $a \in A$ (или любой функции $f \in \mathcal{F}$ из соответствующего ему класса эквивалентности) на произвольной выборке $X \subseteq \mathbb{X}$ определяется как

$$n(a, X) = \operatorname{card} \{x \in X\colon I(a, x) = 1\} .$$

Частота ошибок определяется как $\nu(a, X) = \frac{n(a,X)}{|X|}$. Будем называть частоту ошибок на полной выборке $\nu(a, \mathbb{X})$ *истинной частотой ошибок* алгоритма a; число ошибок на полной выборке $n(a, \mathbb{X})$ — *полным числом ошибок*. Частота ошибок алгоритма a на полной выборке $\mathbb{X}$ представляет аналог вероятности ошибки алгоритма в стандартном PAC-подходе.

В дальнейшем качество алгоритмов будет характеризоваться только числом или частотой ошибок — величинами, зависящими от бинарного вектора ошибок. Поэтому алгоритмы с одинаковыми векторами ошибок можно считать неразличимыми и рассматривать *метод обучения* μ как отображение $\mu\colon 2^{\mathbb{X}} \to A$.

Обозначим через $[\mathbb{X}]^\ell$ множество всех $\binom{L}{\ell}$ подмножеств из $\mathbb{X}$ мощности ℓ:

$$[\mathbb{X}]^\ell = \{X \subset \mathbb{X}\colon |X| = \ell\} .$$

Будем называть подмножества $X \in [\mathbb{X}]^\ell$ *обучающими* или *наблюдаемыми* выборками, $\nu(a, X)$ — *наблюдаемой частотой ошибок*, $n(a, X)$ — *наблюдаемым числом*

ошибок. Будем также пользоваться сокращенными обозначениями

$$n(a,\mathbb{X}) \equiv n_a, \quad n(a,X) \equiv \hat{n}_a, \quad \nu(a,\mathbb{X}) \equiv \nu_a, \quad \nu(a,X) \equiv \hat{\nu}_a.$$

Будем обозначать через

$$\nu \in \left\{\tfrac{0}{L}, \ldots, \tfrac{L}{L}\right\}, \qquad \hat{\nu} \in \left\{\tfrac{0}{\ell}, \ldots, \tfrac{\ell}{\ell}\right\}$$

(без индекса a) — допустимые значения частоты на полной/обучающей выборке и

$$m \in \{0, \ldots, L\}, \qquad s \in \{0, \ldots, \ell\}$$

— допустимые значения числа ошибок на полной/обучающей выборке.

Допустим, что все разбиения полной выборки $\mathbb{X}$ на две подвыборки, наблюдаемую обучающую X длины ℓ и скрытую контрольную $\mathbb{X}\backslash X$ длины $L-\ell$, равновероятны. Это предположение является ослабленным вариантом стандартной гипотезы о независимости объектов выборки при выборе из распределения вероятностей. Будем называть произвольный предикат $\varphi\colon [\mathbb{X}]^\ell \to \{\text{истина, ложь}\}$ *событием*. Определим *вероятность события* φ как долю выборок в $[\mathbb{X}]^\ell$, для которых φ истинен:

$$\mathbf{P}[\,\varphi(X)] \stackrel{\text{def}}{=} \frac{1}{\binom{L}{\ell}} \sum_{X \in [\mathbb{X}]^\ell} [\,\varphi(X)]\,.$$

Для произвольных предикатов φ_1, φ_2 будем обозначать через $\varphi_1 \vee \varphi_2$, $\varphi_1\varphi_2$, $\bar{\varphi}_1$ их дизъюнкцию, конъюнкцию и отрицание, соответственно. Для набора предикатов $\varphi_1, \ldots, \varphi_k$ будем обозначать дизъюнкцию и конъюнкцию как $\vee_i \varphi_i$ и $\wedge_i \varphi_i$. Отметим, что $\mathbf{P}\big[\wedge_{i=1}^k \varphi_i\big] \equiv \mathbf{P}\,[\varphi_1] \ldots [\varphi_k]$.

Произвольную функцию $\xi\colon [\mathbb{X}]^\ell \to Z$, где Z — некоторое множество, будем называть случайной величиной. В случае $Z = \mathbb{R}$ определим математическое ожидание как среднее значение по всем разбиениям:

$$\mathbf{E}\xi(X) \stackrel{\text{def}}{=} \binom{L}{\ell}^{-1} \sum_X \xi(X).$$

Алгоритм

$$\hat{a} \stackrel{\text{def}}{=} \mu(X),$$

получаемый в результате обучения, является случайной величиной, принимающей значения из A.

Нашей основной задачей является получение как можно более точных доверительных оценок $\bar{\nu}$ истинной частоты ошибок:

$$\mathbf{P}[\,\nu_{\hat{a}} < \bar{\nu}(\hat{a}, X, \mathcal{F}, \mu, \alpha)] \;\geqslant\; 1-\alpha,$$

где $\bar{\nu}$ некоторая оценочная функция и $\alpha \in (0,1)$ достаточно близко к нулю. Назначение таких оценок состоит в том, чтобы, минимизируя оценочную функцию $\bar{\nu}(\hat{a}, X, \mathcal{F}, \mu, \alpha)$ по $\mathcal{F}, \mu$, выбрать метод обучения μ и семейство $\mathcal{F}$ с наилучшей обобщающей способностью.

1.2 Вероятность перебучения и обращение оценок

Доверительные оценки частоты ошибок можно получать путём обращения оценок вероятности переобучения.

Обычно говорят, что алгоритм $a \in A$ является *переобученным* на обучающей выборке $X \in [\mathbb{X}]^\ell$, если его частота ошибок на X существенно меньше его частоты ошибок на полной выборке $\mathbb{X}$, то есть если $\nu_a - \hat{\nu}_a \geqslant \varepsilon$, где *порог переобучения* ε принимает значения из $[-1, +1]$.

В теории статистического обучения условие переобучения также часто определяется через частоту ошибок на контрольной выборке: $\nu(a, \bar{X}) - \nu(a, X) \geqslant \varepsilon$. Легко проверить, что оно эквивалентно условию $\nu_a - \hat{\nu}_a \geqslant \varepsilon$ с точностью до сомножителя: $\nu_a - \hat{\nu}_a \geqslant \frac{k}{L}\varepsilon$.

Определим обобщенное условие переобучения

Определение 1.2.1. *Критерием переобучения назовем предикат*

$$\mathrm{U}_a \colon [\mathbb{X}]^\ell \to \{\textit{истина, ложь}\}$$

вида

$$U(n_a, \hat{n}_a) \geqslant \varepsilon, \tag{1.2}$$

где $n_a \in \{0, \ldots, L\}$, $\hat{n}_a \in \{0, \ldots, \ell\}$ *и* $U(n_a, \hat{n}_a)$ *— некоторая функция, неубывающая по* n_a *и невозрастающая по* $\hat{n}_a$

Утверждение 1.2.2. *Критерий переобучения* (1.2) *равносильно записывается как*

$$n_a \geqslant \bar{n}\,(\hat{n}_a, \varepsilon), \qquad \textit{где} \qquad \bar{n}\,(\hat{n}_a, \varepsilon) = \min\{n \colon U(n, \hat{n}_a) \geqslant \varepsilon\} \tag{1.3}$$

и как

$$\hat{n}_a \leqslant s_{n_a}(\varepsilon), \qquad \textit{где} \qquad s_{n_a}(\varepsilon) = \max\{\hat{n} \colon U(n_a, \hat{n}) \geqslant \varepsilon\}\,. \tag{1.4}$$

Для удобства доопределим функции $\bar{n}\,(\hat{n}_a, \varepsilon)$ и $s_{n_a}(\varepsilon)$ соответственно как $L+1$ и -1 для экстремальных значений их параметров при которых множества в их определениях пусты.

Утверждение 1.2.3. *Функция $\bar{n}\,(\hat{n}_a, \varepsilon)$ не убывает по $\hat{n}_a$ и не убывает по ε. Функция $s_{n_a}(\varepsilon)$ не убывает по n_a и не возрастает по ε.*

Определение 1.2.4. *Областью переобучения будем называть множество значений*

$$\{(n_a, \hat{n}_a)\colon U(n_a, \hat{n}_a) \geqslant \varepsilon\}\,.$$

Границей переобучения будем называть s_{n_a} как функцию параметра n_a.

Пример области переобучения и границы переобучения для критерия $\nu_a - \hat{\nu}_a \geqslant \varepsilon$ приведен на Рис. 1.1.

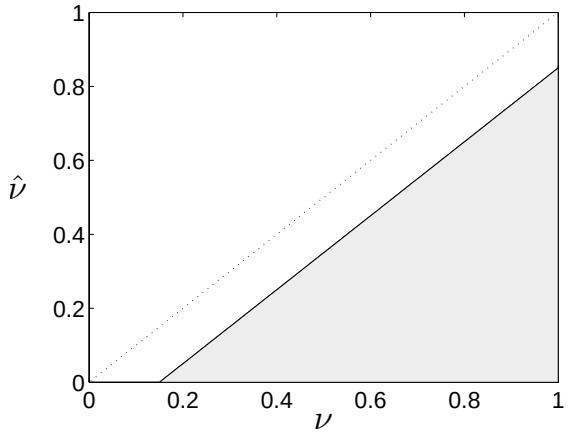

Рис. 1.1. Область переобучения и граница переобучения для критерия $\nu_a - \hat{\nu}_a \geqslant \varepsilon$ ($\varepsilon = 15\%$). Диагональ $\hat{\nu}_a = \nu_a$ соответствует нулевому уклонению частот; серая область соответствует парам $(\nu_a, \hat{\nu}_a)$, удовлетворяющим условию переобучения, граница серой области соответствует парам $(n_a, s_{n_a}(\varepsilon))$, $n_a = 1, \ldots, L$ или парам $(\bar{n}\,(\hat{n}_a, \varepsilon), \hat{n}_a)$, $\hat{n}_a = 1, \ldots, \ell$.

В дальнейшем ограничимся рассмотрением достаточно «гладких» границ переобучения, для которых $s_{m+1} - s_m \in \{0, 1\}$. Это очевидно выполняется для критериев $\nu_a - \hat{\nu}_a \geqslant \varepsilon$ и $\nu(a, \bar{X}) - \nu(a, X) \geqslant \varepsilon$.

Будем говорить, что *метод μ переобучается* на выборке X, если он выбирает из A переобученный алгоритм, то есть если для X выполняется $U(n_{\hat{a}}, \hat{n}_{\hat{a}}) \geqslant \varepsilon$.

Определение 1.2.5. *Вероятность переобучения метода μ есть*

$$\mathbf{P}[\,U(n_{\hat{a}}, \hat{n}_{\hat{a}}) \geqslant \varepsilon]\,,$$

она равна доле выборок $X \in [\mathbb{X}]^\ell$, для которых алгоритм $\hat{a} = \mu(X)$ переобучен.

Стандартному критерию переобучения $\nu - \hat{\nu} \geqslant \varepsilon$ соответствует функционал вероятности переобучения

$$Q_\varepsilon \stackrel{\text{def}}{=} \mathbf{P}[\,\nu_{\hat{a}} - \hat{\nu}_{\hat{a}} \geqslant \varepsilon]\,.$$

Утверждение 1.2.6. *Если имеется верхняя оценка вероятности переобучения*

$$\mathbf{P}[\,U(n_{\hat{a}}, \hat{n}_{\hat{a}}) \geqslant \varepsilon] \leqslant P(\varepsilon), \tag{1.5}$$

где $P(\varepsilon)$ — невозрастающая функция от ε, то определяя

$$P^{-1}(\alpha) = \min\{\varepsilon\colon P(\varepsilon) \leqslant \alpha\}, \quad \alpha \in (0,1),$$

имеем доверительную оценку

$$\mathbf{P}\big[\, n_{\hat{a}} \geqslant \bar{n}\left(\hat{n}_{\hat{a}}, P^{-1}(\alpha)\right)\big] \leqslant \alpha. \tag{1.6}$$

Последняя оценка означает, что с вероятностью не менее $1-\alpha$ полное число ошибок $n_{\hat{a}}$ алгоритма $\hat{a} = \mu(X)$ не превышает $\bar{n}\,(\hat{n}_{\hat{a}}, P^{-1}(\alpha))$. Будем называть переход от (1.5) к (1.6) *обращением* оценки вероятности переобучения.

1.3 Гипергеометрическое распределение

Пусть имеется L объектов, из которых равновероятно выбирается без возвращения ℓ объектов. Если среди L объектов m «помечены», то вероятность того, что в выборку попадут s помеченных объектов, подчиняется гипергеометрическому распределению [39]:

$$h_{L,m}^{\ell}(s) \stackrel{\text{def}}{=} \frac{\binom{m}{s}\binom{L-m}{\ell-s}}{\binom{L}{\ell}} \tag{1.7}$$

— унимодальному несимметричному распределению со средним значением $\mathbf{E}s = m\frac{\ell}{L}$. Область определения $s = \max\{0, m+\ell-L\}, \ldots, \min\{m, \ell\}$. Функцию распределения будем обозначать

$$H_{L,m}^{\ell}(s) \stackrel{\text{def}}{=} \sum_{t=0}^{s} h_{L,m}^{\ell}(t)\,.$$

Далее будем опускать для краткости индексы L, ℓ и писать $h_m(s)$ and $H_m(s)$. Условимся, что вне области определения по параметрам L, ℓ, m, s значение биномиальных коэффициентов и гипергеометрической вероятности равно нулю.

Гипергеометрическое распределение (1.7) служит аналогом биномиального распределения в стандартном PAC-подходе; биномиальное распределение есть аппроксимация гипергеометрического при $L \to \infty,\ m/L \to p$.

Отметим, что:

Утверждение 1.3.1. *Внутри своей области определения функция $H_m(s)$ строго возрастает по s и строго убывает по m.*

Утверждение 1.3.2. *При $\frac{s}{\ell} < \frac{m}{L}$ (то есть в левом «хвосте» распределения) функция $h_m(s)$ строго возрастает по s и строго убывает по m.*

Утверждение 1.3.3. *Для гипергеометрического распределения выполняется соотношение $H_m(s) - H_{m-1}(s-1) = \frac{m-s}{m} h_m(s)$.*

Доказательство. Рассмотрим производящую функцию гипергеометрического распределения $\mathcal{R} = (1+zx)^m (1+z)^{L-m}$. Имеем $\left[z^\ell x^{\leqslant s} \right] \mathcal{R} = C_L^\ell H_m(s)$ (здесь $\left[z^\ell x^{\leqslant s} \right]$ есть обозначение оператора дающего сумму коэффициентов при одночленах степени ℓ по z и не более s по x).

«Отметим» один из объектов и рассмотрим функцию $\tilde{\mathcal{R}} = (1+zx)^{m-1}(1+zxy)(1+z)^{L-m}$. По-прежнему имеем $\left[z^\ell x^{\leqslant s} \right] \tilde{\mathcal{R}} = H_m(s)\, C_L^\ell$, с другой стороны, раскладывая на слагаемые имеем:

$$\left[z^\ell x^{\leqslant s} \right] \tilde{\mathcal{R}} = \left[z^\ell x^{\leqslant s} \right] \sum_{i=0}^{m-1} \sum_{j=0}^{L-m} C_{m-1}^i C_{L-m}^j z^{i+j} x^i \cdot (1 + zxy) =$$

$$= \sum_{i=0}^{s} C_{m-1}^i C_{L-m}^{\ell-i} + \sum_{i=0}^{s-1} C_{m-1}^i C_{L-m}^{\ell-i-1} = \sum_{i=0}^{s-1} C_{m-1}^i \left(C_{L-m}^{\ell-i} + C_{L-m}^{\ell-i-1} \right) + C_{m-1}^s C_{L-m}^{\ell-s} =$$

$$= \sum_{i=0}^{s-1} C_{m-1}^i C_{L-m+1}^{\ell-i} + C_{m-1}^s C_{L-m}^{\ell-s} = H_{m-1}(s-1)\, C_L^\ell + C_{m-1}^s C_{L-m}^{\ell-s}$$

Используя теперь $\frac{C_{m-1}^s C_{L-m}^{\ell-s}}{C_L^\ell} = \frac{m-s}{m} h_m(s)$, получаем доказываемое утверждение. ■

Следствие 1. *Используя рекуррентные соотношения для гипергеометрической вероятности $h_m(s)$ несложно получить из последнего утверждения следующие соотношения:*

$$\begin{array}{lcllcl} H_m(s) & - & H_{m+1}(s) & = & \frac{\ell-s}{L-m} h_m(s) & = & \frac{s+1}{m+1} h_{m+1}(s+1) \\ H_{m+1}(s+1) & - & H_m(s) & = & \frac{\ell-s}{L-m} \frac{m-s}{s+1} h_m(s) & = & \frac{m-s}{m+1} h_{m+1}(s+1) \end{array}$$

1.4 Оценка для одного алгоритма

Рассмотрим фиксированный алгоритм $a \in A$ с полным числом ошибок $n_a = m$. Число ошибок $\hat{n}_a$ алгоритма a на обучающей выборке подчиняется гипергеометрическому распределению:

$$\begin{aligned} \mathbf{P}[\,\hat{n}_a = s] &= h_m(s)\,; \\ \mathbf{P}[\,\hat{n}_a \leqslant s] &= H_m(s)\,. \end{aligned}$$

1.4.1 Стандартный критерий переобучения

Условие переобучения $\nu_a - \hat{\nu}_a \geqslant \varepsilon$ эквивалентно записывается как

$$\hat{n}_a \leqslant \tfrac{\ell}{L} n_a - \varepsilon\ell = s_{n_a}(\varepsilon).$$

То есть, алгоритм a переобучен, если число его ошибок на обучающей выборке меньше порога s_{n_a}, зависящего от его полного числа ошибок.
Вероятность того, что алгоритм a переобучен:

$$\mathbf{P}[\,\nu_a - \hat{\nu}_a \geqslant \varepsilon] \;=\; \mathbf{P}[\,\hat{n}_a \leqslant s_{n_a}(\varepsilon)] \;=\; H_{n_a}(s_{n_a}(\varepsilon))\,, \tag{1.8}$$

ее верхняя оценка

$$\mathbf{P}[\,\nu_a - \hat{\nu}_a \geqslant \varepsilon] \;\leqslant\; \max_m H_m(s_m(\varepsilon)) \;\stackrel{\text{def}}{=}\; P_1(\varepsilon). \tag{1.9}$$

Отметим, что здесь $H_m(s_m)$ служит точным (не асимптотическим) аналогом верхних оценок «хвоста» биномиального распределения (к примеру, оценок Чернова или Хёффдинга) используемых в стандартном PAC-подходе.

Обращая оценку (1.9), получаем доверительную оценку для ν_a:

$$\forall\alpha \in (0,1) \quad \mathbf{P}[\,\nu_a < \hat{\nu}_a + \varepsilon_1(\alpha)] \geqslant 1-\alpha, \quad \text{где} \quad \varepsilon_1(\alpha) = \min\{\varepsilon\colon P_1(\varepsilon) \leqslant \alpha\}\,.$$

1.4.2 Квантильный критерий переобучения

Можно получить более точную доверительную оценку для одного алгоритма, избавившись от огрубления $\max\limits_m$ в (1.9). Идея состоит в том, чтобы использовать вместо разности $\nu_a - \hat{\nu}_a$ другую меру уклонения $\hat{\nu}_a$ от ν_a, распределение которой не зависит от n_a как (1.8). В качестве такой меры уклонения можно взять величину $H_{n_a}(\hat{n}_a)$, определив критерий переобучения как $H_{n_a}(\hat{n}_a) \leqslant \eta$.

Лемма 1.4.1. *Для любого алгоритма $a \in A$ и любого $\eta \in (0,1)$ справедлива доверительная оценка:*

$$\mathbf{P}[\,n_a < \bar{n}\,(\hat{n}_a, \eta)] \;\geqslant\; 1-\eta,$$

где $\bar{n}\,(\hat{n}_a, \eta) = \min\{m\colon H_m(\hat{n}_a) \leqslant \eta\}$.

Доказательство. Поскольку $H_m(\cdot)$ — функция распределения, для любого $\eta \in (0,1)$

$$\mathbf{P}[\,H_{n_a}(\hat{n}_a) \leqslant \eta] = \mathbf{P}[\,1 - H_{n_a}(\hat{n}_a) \geqslant 1-\eta] \leqslant \eta.$$

Используя монотонность $H_m(s)$ (Утверждение 1.3.1), положим в Утверждении 1.2.2 $U(n_a, \hat{n}_a) \equiv 1 - H_{n_a}(\hat{n}_a)$ и $\varepsilon \equiv 1-\eta$. Обращая последнюю оценку в соответствии с Утверждением 1.2.6, имеем утверждение леммы.

■

Из Утверждения 1.2.2 следует, что критерий переобучения $H_{n_a}(\hat{n}_a) \leqslant \eta$ может быть записан в трёх эквивалентных видах:

$$H_{n_a}(\hat{n}_a) \leqslant \eta \quad \Leftrightarrow \quad n_a \geqslant \bar{n}\,(\hat{n}_a, \eta) \quad \Leftrightarrow \quad \hat{n}_a \leqslant s_{n_a}(\eta), \tag{1.10}$$

где

$$\bar{n}\,(\hat{n}_a, \eta) = \min\{m\colon H_m(\hat{n}_a) \leqslant \eta\} \tag{1.11}$$

и

$$s_{n_a}(\eta) = \max\{s\colon H_{n_a}(s) \leqslant \eta\}. \tag{1.12}$$

Третий вариант интерпретируется следующим образом: алгоритм a переобучен, если число его ошибок $\hat{n}_a$ на обучающей выборке меньше η-квантили распределения $H_{n_a}(s)$. Предполагается, что значение параметра η достаточно близко к нулю, так что $\hat{n}_a$ попадает в левый хвост гипергеометрического распределения (при этом $\hat{\nu}_a < \nu_a$). Будем называть условие $H_{n_a}(\hat{n}_a) \leqslant \eta$ *квантильным критерием переобучения*. Квантильный критерий удобен тем, что вероятность переобучения для одного алгоритма равна η независимо от его полного числа ошибок n_a.

Для простоты, везде далее будем далее использовать функции $\bar{n}$ и s_{n_a} в виде (1.11), (1.12) и, соответственно, функционал вероятности переобучения в виде:

$$Q_\eta \stackrel{\text{def}}{=} \mathbf{P}[\,H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta\,].$$

Отметим, что оценки всех последующих теорем с небольшими модификациями могут быть записаны для произвольного критерия.

Параметр η здесь играет ту же роль, что ранее играл ε, и мы его также будем называть *порогом переобучения*. Заметим однако, что чем больше η, тем меньше уклонение частот ошибок, и тем *менее* переобучен алгоритм a, для ε зависимость обратная: чем больше ε, тем *более* переобучен алгоритм a.

На Рис. 1.2 для обоих условий переобучения серым фоном отмечены удовлетворяющие им области значений ν_a и $\hat{\nu}_a$. Квантильное условие переобучения дает более точные доверительные оценки для алгоритмов с малыми и больши́ми значениями частоты ν_a, т. к. оно учитывает, что дисперсия гипергеометрического распределения в этих случаях меньше, чем при $\nu_a \approx \frac{1}{2}$.

Определим верхнюю оценку истинной частоты ошибок алгоритма

$$\bar{\nu}\,(\hat{\nu}_a, \eta) \stackrel{\text{def}}{=} \bar{n}\,(\hat{\nu}_a \ell, \eta)/L.$$

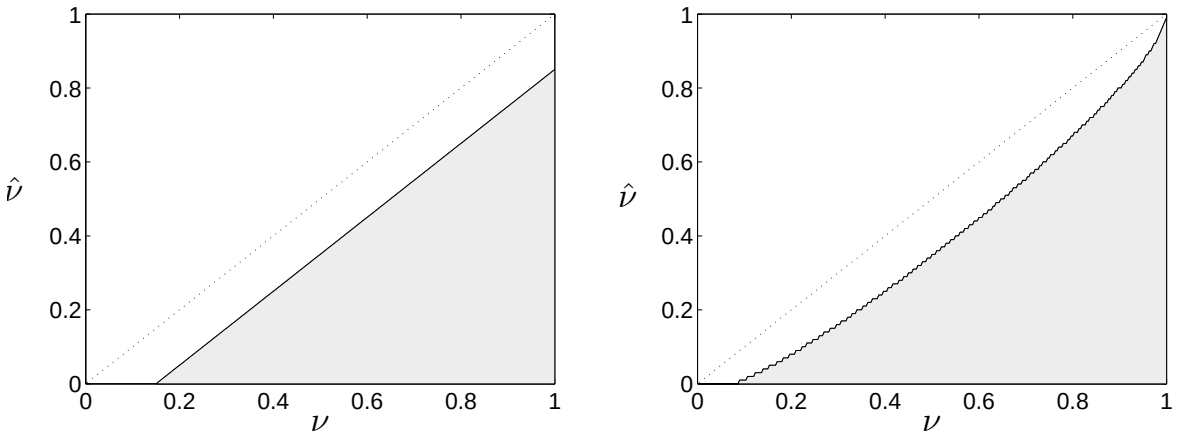

Рис. 1.2. Стандартное условие переобучения $\nu_a - \hat{\nu}_a \geqslant \varepsilon$ (слева, $\varepsilon = 15\%$, $P_1(\varepsilon) \approx 2 \cdot 10^{-4}$) и квантильное условие переобучения $H_{L\nu_a}(\ell\hat{\nu}_a) \leqslant \eta$ (справа, $\eta = P_1(\varepsilon)$), при $L = 300, \ell = 100$. Диагональ $\hat{\nu}_a = \nu_a$ соответствует нулевому уклонению частот; серая область соответствует парам $(\nu_a, \hat{\nu}_a)$, удовлетворяющим условию переобучения. Горизонтальное расстояние от диагонали до области переобучения есть $(\bar{\nu}\,(\hat{\nu}_a, \eta) - \hat{\nu}_a)$ — разница между верхней оценкой истинной частоты по наблюдаемой частоте и наблюдаемой частотой.

В дальнейшем нам потребуется также нижняя доверительная оценка для n_a. Аналогично Лемме 1.4.1 имеем

Лемма 1.4.2. *Для любого алгоритма $a \in A$ и любого $\eta \in (0, 1)$ выполняется доверительная оценка*

$$\mathbf{P}[\, n_a > \underline{n}\,(\hat{n}_a, \eta)] \;\geqslant\; 1 - \eta,$$

где $\underline{n}\,(\hat{n}_a, \eta) = \max\{m \colon 1 - H_m(\hat{n}_a) \leqslant \eta\}$.

Доказательство. Очевидно, $\mathbf{P}[\, 1 - H_{n_a}(\hat{n}_a) \leqslant \eta] \leqslant \eta$. Поскольку $1 - H_{n_a}(\hat{n}_a)$ возрастает по n_a, имеем $1 - H_{n_a}(\hat{n}_a) \leqslant \eta \;\Leftrightarrow\; n_a \leqslant \underline{n}\,(\hat{n}_a, \eta)$. ■

1.5 Комбинаторные оценки Вапника-Червоненкиса и «бритвы Оккама»

Чтобы получить верхнюю оценку вероятности переобучения $\mathbf{P}\Big[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta\Big]$, справедливую для любого метода обучения μ, в теории Вапника-Червоненкиса вводится *принцип равномерной сходимости* [65]. Вероятность переобучения метода μ оценивается сверху вероятностью того, что в A *можно* выбрать переобученный алгоритм (шаг I в (1.13)). Затем применяется неравенство Буля (шаг II в (1.13)).

Теорема 1.5.1 (VC-оценка). *Для любого семейства $\mathcal{F}$, любой полной выборки $\mathbb{X}, |\mathbb{X}| = L$, метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$, индикатора ошибки $\mathrm{I}\colon \mathbb{X} \times \mathcal{F} \to \{0,1\}$, $\forall \alpha \in (0,1)$ верна оценка вероятности переобучения*

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta)] \;\leqslant\; P_{\mathrm{VC}}(\eta) \;\stackrel{\text{def}}{=}\; |A|\,\eta$$

и доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{\mathrm{VC}}(\alpha))] \;\leqslant\; \alpha,$$

где $\eta_{\mathrm{VC}}(\alpha) = \frac{\alpha}{|A|}$.

Доказательство. Вероятность переобучения метода μ ($\hat{a} = \mu X$) оценивается как

$$\mathbf{P}\Big[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta\Big] \;\stackrel{\mathrm{I}}{\leqslant}\; \mathbf{P}\Big[\, \exists a \in A\colon H_{n_a}(\hat{n}_a) \leqslant \eta\Big] \;\stackrel{\mathrm{II}}{\leqslant}\; \sum_{a \in A} \mathbf{P}[\, H_{n_a}(\hat{n}_a) \leqslant \eta] = |A|\,\eta. \quad (1.13)$$

Фиксируя доверительную вероятность $|A|\,\eta = \alpha$, и обращая условие переобучения $H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta$ в соответствии с Утверждением 1.2.2, имеем утверждение теоремы. ■

Далее будем для краткости называть оценку последней теоремы VC-оценкой. Отметим, что в этой и последующих оценках величина α имеет смысл вероятности переобучения, а величина η — порога переобучения. Для случая одного алгоритма они совпадают по величине.

На практике неравенство Буля оказывается сильно завышенным, поскольку число событий вида $H_{n_a}(\hat{n}_a) \leqslant \eta,\ a \in A$, очень велико, и они существенно совместны. Число этих событий $|A|$, называемое в теории Вапника-Червоненкиса *коэффициентом разнообразия*, может быть порядка 10^{10}–10^{20} (к примеру, для $L = 200 \div 300$ и семейства $\mathcal{F}$ линейных классификаторов). В результате VC-оценка вероятности переобучения оказывается завышенной на несколько порядков.

Оценка «бритвы Оккама» [55, 47], хотя и не устраняет факторов завышенности VC-оценки, позволяет добиться количественного улучшения за счет использования априорной информации о том, что какие-то алгоритмы получаются в результате обучения чаще других. Идея заключается в том, чтобы позволить порогу η зависеть от алгоритма a и априори задать «жесткость» условия переобучения индивидуально для каждого алгоритма $a \in A$. Тогда вероятность переобучения определяется функционалом

$$Q_{occam} \;\stackrel{\text{def}}{=}\; \mathbf{P}\big[H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta_{\hat{a}}\big],$$

зависящим уже от *набора* порогов $\boldsymbol{\eta} = \{\eta_a\colon a \in A\}$, и может быть оценена сверху аналогично предыдущей теореме:

Теорема 1.5.2 (оценка «бритвы Оккама»). *Для любого семейства $\mathcal{F}$, любой полной выборки $\mathbb{X}, |\mathbb{X}| = L$, метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$, индикатора ошибки $\mathrm{I}\colon \mathbb{X} \times \times \mathcal{F} \to \{0,1\}$, $\forall \alpha \in (0,1)$ и вектора порогов переобучения $\boldsymbol{\eta} = \{\eta_a\colon a \in A\}$ верна оценка вероятности переобучения*

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{\hat{a}})] \;\leqslant\; \sum_{a \in A} \eta_a.$$

При условии $\sum_{a\in A} \eta_a = \alpha$, верна также доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{\hat{a}})] \;\leqslant\; \alpha. \tag{1.14}$$

Доказательство.

$$\mathbf{P}\Big[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta_{\hat{a}}\Big] \leqslant \mathbf{P}\Big[\, \exists a \in A\colon H_{n_a}(\hat{n}_a) \leqslant \eta_a)\Big] \leqslant \sum_{a\in A} \mathbf{P}[\, H_{n_a}(\hat{n}_a) \leqslant \eta_a)] = \sum_{a\in A} \eta_a. \tag{1.15}$$

Обращая условие переобучения имеем утверждение теоремы. ■

Заметим, что если среднее значение порогов η_a по множеству A есть η, то имеем $\sum_{a\in A} \eta_a = \eta\,|A|$, и в теореме имеем оценку вероятности переобучения в точности равную VC-оценке. Однако при этом оценка теоремы позволяет получить *более точные в среднем (по обучающим выборкам) доверительные оценки* для $n_{\hat{a}}$, задавая бóльшие пороги η_a (и, соответственно, получая меньшие оценки $\bar{n}\,(\hat{n}_a, \eta_a)$) для тех a, которые чаще являются результатом обучения.

В вырожденном случае, когда при обучении методом μ из A всегда выбирается один и тот же алгоритм a_0, мы можем исключить из оценки все остальные алгоритмы, положив $\eta_a = 0$ для $a \neq a_0$ и $\eta_{a_0} = \alpha$. Тогда последняя теорема дает доверительную оценку $n_{\hat{a}} \leqslant \bar{n}\,(\hat{n}_{\hat{a}}, \alpha)$ совпадающую с оценкой для одного алгоритма из Леммы 1.4.1, в то время как Теорема 1.5.1 по прежнему дает $n_{\hat{a}} \leqslant \bar{n}\,(\hat{n}_{\hat{a}}, \alpha/|A|)$.

1.6 Оптимальный набор весов в оценке Occam razor

Определим в некотором смысле оптимальные значения порогов в оценке Теоремы 1.5.2. Пусть задан произвольный набор неотрицательных весов $\boldsymbol{q} = (q_a\colon a \in A)$ такой, что $\sum_{a\in A} q_a = 1$ и пусть $\boldsymbol{\eta} = \boldsymbol{q}\alpha$. Очевидно, что за счет подбора весов $\boldsymbol{q}$ можно получить меньшую в среднем доверительную оценку:

$$\min_{\boldsymbol{q}} \mathbf{E}\left(\bar{\nu}\,(\hat{\nu}_{\hat{a}}, q_{\hat{a}}\alpha) - \hat{\nu}_{\hat{a}}\right) \;\leqslant\; \mathbf{E}\left(\bar{\nu}\left(\hat{\nu}_{\hat{a}}, \tfrac{\alpha}{|A|}\right) - \hat{\nu}_{\hat{a}}\right).$$

Здесь разность $\bar{\nu}\,(\hat{\nu}_{\hat{a}}, \eta_{\hat{a}}) - \hat{\nu}_{\hat{a}}$ имеет смысл верхней оценки уклонения истинной частоты от наблюдаемой $\nu_{\hat{a}} - \hat{\nu}_{\hat{a}}$. Мы минимизируем разность, а не непосредственно оценку $\bar{\nu}\,(\hat{\nu}_{\hat{a}}, \eta_{\hat{a}})$ чтобы исключить зависимость минимума от наблюдаемых частот $\hat{\nu}_{\hat{a}}$, выбираемых методом μ при обучении.

Рассмотрим вместо разности $\nu_a - \hat{\nu}_a$ другую меру уклонения частот — логарифм функции гипергеометрического распределения

$$d(\hat{n}_a, n_a) = -\ln H_{n_a}(\hat{n}_a)\,.$$

Действительно, в силу известной оценки $H_{n_a}(\hat{n}_a) \leqslant e^{-\mathrm{Const}\cdot(\nu_a - \hat{\nu}_a)^2}$ [38], имеем

$$(\nu_a - \hat{\nu}_a)^2 \leqslant \mathrm{Const}\cdot d(\hat{n}_a, n_a)$$

то есть $d(\hat{n}_a, n_a)$ имеет смысл верхней оценки квадрата уклонения частот $(\nu_a - \hat{\nu}_a)^2$.

Далее, заметим, что из определения $\bar{n}\,(\hat{n}_a, \eta_a)$ следует $d(\hat{n}_a, \bar{n}\,(\hat{n}_a, \eta_a)) \approx -\ln \eta_a$. Здесь равенство практически точное; приближённое равенство возникает только вследствие дискретности значений $H_m(s)$. Таким образом,

$$(\bar{\nu}\,(\hat{n}_a, \eta_a) - \hat{\nu}_a)^2 \leqslant \mathrm{Const}\cdot(-\ln \eta_a),$$

то есть $-\ln \eta_a$ имеет смысл величины оценки $(\bar{\nu}\,(\hat{n}_a, \eta_a) - \hat{\nu}_a)^2$ которую мы будем иметь для алгоритма a при выбранном значении порога η_a.

Чтобы понять, насколько эта оценка велика в среднем при заданном выборе порогов $\boldsymbol{\eta}$, оценим математическое ожидание $\mathbf{E}\,(-\ln \eta_{\hat{a}})$.

Лемма 1.6.1. *Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ и метода обучения μ, пусть $\boldsymbol{p} = \Big(\mathbf{P}[\,\mu(X) = a] :\ a \in A\Big)$ есть распределение вероятностей получения различных алгоритмов в результате обучения и $\boldsymbol{\eta} = \boldsymbol{q}\alpha$, $\sum\limits_{a\in A} q_a = 1$. Тогда минимум*

$$\min_{\boldsymbol{q}} \mathbf{E}\,(-\ln \eta_{\hat{a}}) \qquad \textit{при условии} \quad \sum_{a\in A} q_a = 1$$

достигается при $\boldsymbol{q} = \boldsymbol{p}$.

Доказательство. Имеем задачу минимизации:

$$\min_{\boldsymbol{q}} \mathbf{E}\,(-\ln q_{\hat{a}}) = \min_{\boldsymbol{q}} \sum_{a\in A} p_a \ln \frac{1}{q_a}.$$

Добавляя константу $\sum\limits_{a\in A} p_a \ln p_a$, имеем задачу минимизации с минимумом в той же точке:

$$\min_{\boldsymbol{q}} \sum_{a\in A} p_a \ln \frac{p_a}{q_a}.$$

Минимизируемая величина представляет собой дивергенцию Кульбака-Лейблера между распределением $\boldsymbol{p}$ и набором весов $\boldsymbol{q}$. Как известно, её минимум достигается при $\boldsymbol{q} = \boldsymbol{p}$. ■

Последняя лемма говорит, что оценка «бритвы Оккама» дает оптимальную доверительную оценку, если пороги обучения для отдельных алгоритмов равны вероятностям получения этих алгоритмов в результате обучения.

1.7 Точная оценка вероятности переобучения по методу Монте-Карло

Для определения степени завышенности получаемых далее оценок, определим точное значение вероятности переобучения при заданных $\mathbb{X}$, μ, $\mathcal{F}$. Используем квантильный критерий переобучения $H_n(\hat{n}) \leqslant \eta$. Представим вероятность переобучения в виде разложения:

$$\mathbf{P}[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta] \;=\; \sum_{m=0}^{L} \sum_{s=0}^{\ell} [\, H_m(s) \leqslant \eta]\, P(m,s), \tag{1.16}$$

где

$$P(m,s) \;\stackrel{\text{def}}{=}\; \mathbf{P}[\, n_{\hat{a}} = m \;\wedge\; \hat{n}_{\hat{a}} = s] \tag{1.17}$$

есть вероятность того, что при обучении будет выбран алгоритм с s ошибками на обучающей выборке и m ошибками на полной выборке. Суммирование идет по таким парам (m,s), для которых выполняется условие переобучения при заданном пороге η. Определим также для удобства

$$P(\nu,\hat{\nu}) \;\stackrel{\text{def}}{=}\; \mathbf{P}[\, n_{\hat{a}} = \nu L \;\wedge\; \hat{n}_{\hat{a}} = \hat{\nu}\ell\,].$$

Рассматривая (1.16) как оценку вероятности переобучения и обращая ее в соответствии с Утверждением 1.2.6, получаем следующую (точную) доверительную оценку истинной частоты ошибок:

Лемма 1.7.1. *Для любого семейства $\mathcal{F}$, любой полной выборки $\mathbb{X}, |\mathbb{X}| = L$, метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$, индикатора ошибки $\mathrm{I}\colon \mathbb{X} \times \mathcal{F} \to \{0,1\}$, верна оценка вероятности переобучения*

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta] \;=\; P_{exact}(\eta) \;\stackrel{\text{def}}{=}\; \sum_{m,s} \Big[\, H_m(s) \leqslant \eta \Big]\, P(m,s)$$

и доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{exact}(\alpha))] \;\leqslant\; \alpha.$$

где $\eta_{exact}(\alpha) = \max\{\eta\colon P_{exact}(\eta) \leqslant \alpha\}$ – *эффективный порог переобучения.*

Заметим, что доверительная оценка леммы практически точная: вероятность $\mathbf{P}[\, n_{\hat{a}} < \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{exact}(\alpha))]$ практически равна $1-\alpha$, а неравенство $\geqslant$ появляется только вследствие дискретности значений $H_m(s)$.

Распределение $P(m,s)$ неизвестно, но для фиксированных $\mathbb{X},\ \mu,\ \mathcal{F}$ мы можем получить его оценку по методу Монте-Карло. Для этого используем случайное подмножество обучающих выборок $\mathfrak{X} \subset [\mathbb{X}]^\ell$. Тогда

$$\tilde{P}(m,s) = |\mathfrak{X}|^{-1} \operatorname{card}\{X \in \mathfrak{X}\colon \hat{n}_{\hat{a}} = s \ \wedge\ n_{\hat{a}} = m\}\,.$$

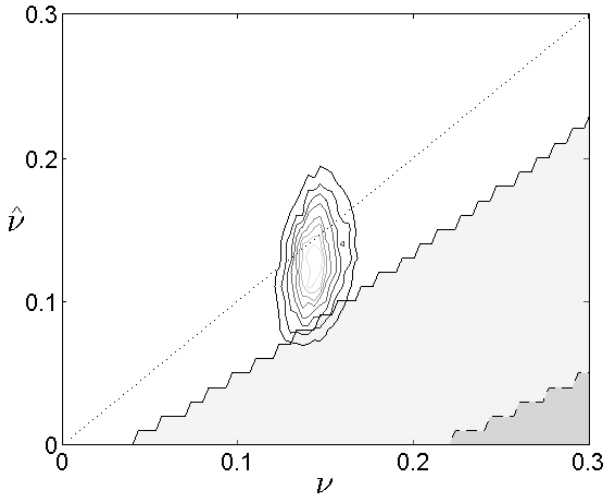

Рис. 1.3. Контурные линии распределения $P(\nu,\hat{\nu})$ для семейства $\mathcal{F}$ линейных классификаторов и метода обучения SVM. Граница переобучения $H_{\nu L}(\hat{\nu}\ell) \leqslant \eta_{exact}(\alpha)$ (сплошная линия) проводится так, чтобы α-квантиль (здесь $\alpha = 10\%$) распределения $P(\nu,\hat{\nu})$ оказалась в зоне переобучения. В данном случае это достигается при пороге переобучения $\eta_{exact}(\alpha) \approx \frac{\alpha}{2.5}$. То есть соответствует VC-оценке с эффективным размером множества A в 2.5 алгоритма. Вторая граница $H_{\nu L}(\hat{\nu}\ell) \leqslant \eta_{\mathrm{VC}}(\alpha)$ (пунктирная линия) соответствует порогу переобучения $\eta_{\mathrm{VC}}(\alpha) = \frac{\alpha}{|A|}$, который дает VC-оценка. Параметры эксперимента $\mathbb{X} \subset \mathbb{R}^5,\ L = 300,\ |A| = 4 \cdot 10^{10},\ |\mathfrak{X}| = 10.000$.

Пример оценки распределения $P(m,s)$ для семейства линейных классификаторов, метода SVM и сгенерированной случайным образом полной выборки $\mathbb{X}$ представлен на Рис. 1.3.

Глава 2

Обзор литературы

В настоящей главе рассматриваются некоторые известные подходы к получению оценок обобщающей способности и решению проблемы переобучения. Большинство описываемых результатов можно отнести к теории статистического обучения (statistical learning theory, SLT). Поскольку все они используют классическую, а не комбинаторную модель вероятности, будем также использовать в настоящей главе классическое определение вероятности.

2.1 Модель обучения

Будем использовать следующую модель обучения. Рассматривается множество $\mathcal{X}$ объектов, множество Y ответов и неизвестное распределение вероятностей $\mathcal{D}$ на $\mathcal{X} \times \times Y$, из которого берется обучающая выборка

$$X = \{(x_i, y_i) \colon i = 1, \ldots, \ell,\ x_i \in \mathcal{X}, y_i \in Y\}.$$

Рассматривается семейство $\mathcal{F} \subset Y^{\mathcal{X}}$ функций из $\mathcal{X}$ в Y и метод обучения $\mu \colon (\mathcal{X} \times \times Y)^{\ell} \to \mathcal{F}$, который выдает некоторую функцию $\hat{f} = \mu(X) \in \mathcal{F}$ в ответ на обучающую выборку X. Пусть $L(f(x), y) \colon Y \times Y \to [0, 1]$ есть некоторая функция потерь, будем обозначать $L_f(x, y) = L(f(x), y)$ потери функции f на $\mathcal{X} \times Y$. Цель метода обучения — выбрать функцию, ожидаемые потери которой $\mathbf{E}L_f$ не намного выше лучших возможных $\inf\{\mathbf{E}L_f \colon f \in \mathcal{F}\}$. На практике ожидаемые потери оцениваются по средним потерям на обучающей выборке

$$\hat{L}_f = \frac{1}{\ell} \sum_{i=1}^{\ell} L_f(x_i, y_i).$$

Будем для краткости обозначать $L_f = \mathbf{E}L_f$ ожидаемые потери функции f. Для бинарной функции потерь $L(y', y) = [\,y' \neq y]$ будем для краткости обозначать $\nu_f =$

$= \mathbf{E}L_f$ — вероятность ошибки функции f и $\hat{\nu}_f = \hat{L}_f$ — частоту ошибок функции f на обучающей выборке.

Нас интересуют оценки вероятности большого уклонения ожидаемых потерь от средних для функций которые выбирает метод μ:

$$\mathbf{P}\Big[\mathbf{E}L_{\hat{f}} - \hat{L}_{\hat{f}} \geqslant \varepsilon\Big] \leqslant P(\varepsilon, \mu, \mathcal{F}, \ell) \tag{2.1}$$

или доверительные оценки этого уклонения при заданном уровне надежности η:

$$\mathbf{P}\Big[\mathbf{E}L_{\hat{f}} - \hat{L}_{\hat{f}} \geqslant \varepsilon(\eta, \mu, \mathcal{F}, X)\Big] \leqslant \eta. \tag{2.2}$$

2.2 Теория Вапника-Червоненкиса

Один из ведущих в настоящее время подходов к анализу обобщающей способности обучаемых алгоритмов, был впервые предложен около 40 лет назад в работах Вапника и Червоненкиса [3] и позднее получил название теории статистического обучения (statistical learning theory). Подход был переоткрыт в англоязычной литературе [63] под названием PAC (probably approximately correct) подхода. В работе [63] больший акцент делается на вычислительной сложности обучения, а развитая на ее основе теория получила название вычислительной теории обучения (computational learning theory).

Комбинаторный аналог одной из основных оценок теории Вапника-Червоненкиса приведен в параграфе 1.5. Основные достоинства этой оценки – независимость от метода обучения μ и вида целевой зависимости $\mathcal{D}$, являются в то же время и ее недостатками – оценка оказывается завышенной, поскольку рассчитана на наихудший вариант целевой зависимости и метода. В то же время, в этих условиях оценка строго говоря неулучшаема – можно построить пример распределения $\mathcal{D}$ для которого оценка точна [23].

Основные факторы завышенности оценок Вапника-Червоненкиса были экспериментально измерены в [4]. Во-первых, это оценивание вероятности равномерной сходимости частоты ошибок к вероятности ошибки по всему семейству алгоритмов, а не вероятности переобучения метода, то есть пренебрежение тем фактом, что метод обучения использует лишь часть семейства. Во-вторых, это применение неравенства Буля, то есть пренебрежение сходством алгоритмов в семействе.

Несложным обобщением VC-оценки является оценка Occam Razor [47], ее комбинаторный аналог также приведен в разделе 1.5.

2.3 Оценки концентрации меры

Оценками степени концентрации случайной величины Z называются оценки вида $\mathbf{P}[\,\mathbf{E}Z - Z > \varepsilon] \leqslant P(\varepsilon)$ или $\mathbf{P}[\,Z > \varepsilon] \leqslant P(\varepsilon)$, где $P(\varepsilon)$ – некоторая невозрастающая по ε функция. Оценки такого рода являются одним из основных инструментов при получении оценок обобщающей способности. Подробные обзоры оценок степени концентрации можно найти в [24], [17].

Заметим, что оценка (2.1) есть по сути оценка степени концентрации случайной величины $Z = \mathbf{E}L_{\hat{f}} - \hat{L}_{\hat{f}}$. Обычно, чтобы исключить зависимость от метода обучения оценивают

$$\mathbf{P}\Big[\,\mathbf{E}L_{\hat{f}} - \hat{L}_{\hat{f}} \geqslant \varepsilon\Big] \;\leqslant\; \mathbf{P}\left[\sup_{f\in\mathcal{F}}\Big(\mathbf{E}L_f - \hat{L}_f\Big) \geqslant \varepsilon\right].$$

Правую часть можно далее оценить сверху при помощи неравенства Буля (для бесконечных семейств $\mathcal{F}$ при этом используется проекция $\mathcal{F}_{\mathbb{X}} = \{(f(x_1),\ldots,f(x_L))\colon f\in\mathcal{F}\}$ семейства $\mathcal{F}$ на конечную выборку $\mathbb{X} = \{x_1,\ldots,x_L\}$), тогда получаем задачу оценивания величины $\mathbf{P}\Big[\,\mathbf{E}L_f - \hat{L}_f \geqslant \varepsilon\Big]$ для *фиксированной* функции f. Очевидно $\mathbf{E}L_f = \mathbf{E}\hat{L}_f$ то есть последнее есть оценка степени концентрации величины $\hat{L}_f$. Величина $\hat{L}_f$ есть сумма независимых случайных величин $L(f(x_i),y_i)$, в связи с этим большую роль в классической SLT играют оценки степени концентрации для суммы случайных величин.

В последующих трех оценках будем полагать функцию f фиксированной и обозначать для краткости

$$\hat{L}_i = L(f(x_i),y_i), \qquad \hat{L} = \frac{1}{\ell}\sum_{i=1}^{\ell}\hat{L}_i.$$

Если $\mathcal{F}$ — функции классификации на 2 класса и $L(f(x),y) = [\,f(x)\neq y]$ индикатор неправильного ответа, то $\mathbf{E}L_f$ есть вероятность события $f(x)\neq y$, $\hat{L}_f$ есть частота этого события в ℓ независимых испытаниях и задача сводится к оцениванию величины левого «хвоста» биномиального распределения $\mathrm{Bin}(\mathbf{E}L_f,\ell)$. Оценкой такого рода является, к примеру, неравенство Чернова:

Теорема 2.3.1 (неравенство Чернова). *Если $\hat{L}_i$ есть бинарные случайные величины, то $\forall\varepsilon>0$:*

$$\mathbf{P}\Big[\,\mathbf{E}\hat{L} - \hat{L} \geqslant \varepsilon\Big] \leqslant e^{-2\ell\varepsilon^2}.$$

Если $L(f(x),y)$ некоторая вещественнозначная ограниченная функция, то имеем задачу оценивания степени концентрации для суммы ограниченных случайных величин. Для этого случая существует несколько классических неравенств.

Теорема 2.3.2 (неравенство Хёффдинга, [36]). *Пусть $\hat{L}_i$ есть независимые случайные величины для которых с вероятностью 1 выполняется $\hat{L}_i \in [a_i, b_i]$. Тогда для любого $\varepsilon > 0$ верно:*

$$\mathbf{P}\Big[\,\hat{L} - \mathbf{E}\hat{L} \geqslant \varepsilon\Big] \leqslant e^{-2\varepsilon^2/\ell^2 \sum_i (b_i - a_i)^2},$$

$$\mathbf{P}\Big[\,\mathbf{E}\hat{L} - \hat{L} \geqslant \varepsilon\Big] \leqslant e^{-2\varepsilon^2/\ell^2 \sum_i (b_i - a_i)^2}.$$

Следующее неравенство использует не только ограниченность переменных $\hat{L}_i$, но и знание об их дисперсии [1].

Теорема 2.3.3 (неравенство Бернштейна). *Пусть $Z_i = \hat{L}_i - \mathbf{E}\hat{L}_i$ есть независимые случайные величины и $Z_i < 1$ с вероятностью 1. Пусть $\sigma^2 = \frac{1}{\ell}\sum_i \mathbf{Var} Z_i$. Тогда для любого $\varepsilon > 0$:*

$$\mathbf{P}\Big[\,\mathbf{E}\hat{L} - \hat{L} \geqslant \varepsilon\Big] \leqslant e^{-\ell\varepsilon^2/2(\sigma^2 + \varepsilon/3)}$$

Для того, чтобы оценить непосредственно степень концентрации случайной величины $Z = \sup\limits_{f \in \mathcal{F}}(\mathbf{E}\hat{L}_f - \hat{L}_f)$ без использования неравенства Буля для перехода к фиксированной функции f, необходимы оценки степени концентрации уже не для суммы случайных величин как выше, а для произвольной функции случайных величин. Одна из наиболее известных оценок такого рода — неравенство ограниченных разностей МакДиармида [52].

Теорема 2.3.4. *Пусть $z_i = (x_i, y_i) \in \mathcal{X} \times Y, i = 1, \ldots, \ell$ независимые случайные величины и $g(z_1, \ldots, z_\ell)$ вещественнозначная функция удовлетворяющая условию «ограниченных разностей», то есть такая, что существуют константы $c_i \in \mathbb{R}$, такие что для любого отдельного аргумента верно*

$$|g(z_1, \ldots, z_i, \ldots, z_\ell) - g(z_1, \ldots, z_i', \ldots, z_\ell)| \leqslant c_i$$

при любых значениях аргументов. Тогда $\forall \varepsilon > 0$:

$$\mathbf{P}[\,|\mathbf{E}g - g| \geqslant \varepsilon\,] \;\leqslant\; 2e^{-2\varepsilon^2/\sum_i c_i^2}.$$

Следствие 2 ([17]). *Пусть функции L_f принимают значения в интервале $[a, b]$ и*

$$g = \sup_{f \in \mathcal{F}} \big|\mathbf{E}L_f - \hat{L}_f\big|.$$

Тогда $\forall \delta > 0$ с вероятностью не менее $1 - \delta$ выполняется:

$$g \leqslant \mathbf{E}g + (b - a)\sqrt{\frac{\log(2/\delta)}{2\ell}}.$$

Неравенство МакДиармида используется подобным образом при доказательстве оценки обобщающей способности основанной на Радемахеровской сложности семейства $\mathcal{F}$. Мы приведем эту оценку в одном из последующих параграфов.

2.4 ε-покрытия семейства алгоритмов

Классическая VC-оценка для бесконечного семейства $\mathcal{F}$ бинарных классификаторов получается за счет перехода к конечной проекции $\mathcal{F}_{\mathbb{X}}$ этого семейства на конечную «составную выборку» (double sample) $\mathbb{X}$. Другой способ свести оценку к конечному случаю – рассмотреть ε-покрытие (covering) $\mathcal{F}$ — понятие, введенное в [43], впервые использованное в теории обучения в работах Вапника-Червоненкиса [2] и детально исследованное во многих поздних работах.

Определение 2.4.1. *Если расстояние между функциями $f, f' \in \mathcal{F}$ есть $\rho(f, f') = \operatorname{card}\{x \in X\colon f(x) \neq f'(x)\}$ и $B_\varepsilon(f) = \{f' \in \mathcal{F}\colon \rho(f, f') \leqslant \varepsilon\}$ есть шар радиуса ε, то множество функций $\mathcal{F}_\varepsilon = \{f_1, \ldots, f_s\}$ называется ε-покрытием $\mathcal{F}$ тогда и только тогда, когда $\mathcal{F} \subset \bigcup_{f \in \mathcal{F}_\varepsilon} B_\varepsilon(f)$.*

Определение 2.4.2. *Минимальная мощность ε-покрытия (covering number) определяется как размер минимального возможного покрытия $N(\mathcal{F}, \varepsilon, X) = \arg\min_{\mathcal{F}_\varepsilon} s$*

Минимальная мощность покрытия представляет меру сложности семейства аналогичную коэффициенту разбиения (shattering coefficient) в VC-оценке. Поскольку ответы алгоритмов из одного ε-шара не сильно отличаются, то за счет огрубления порога переобучения ε (и ухудшения точности оценки) на величину порядка ε, возможно выписать оценку, в которую вместо коэффициента разбиения входит мощность покрытия. Существует множество подобных оценок, одна из них приведена ниже [20].

Теорема 2.4.3.

$$\mathbf{P}\Big[\exists f \in \mathcal{F}\colon \nu_f - \hat{\nu}_f > \varepsilon\Big] \;\leqslant\; 8\mathbf{E}\left(N(\mathcal{F}, \varepsilon, X)\right) e^{-\ell\varepsilon^2/128}.$$

По мере увеличения в последней оценке радиуса покрытия ε, мощность покрытия в правой части падает, что уменьшает оценку, однако, точность оценки в левой части также падает. Таким образом, можно подобрать некоторый оптимальный радиус покрытия, подобная задача решается в цепном методе Дадли [57].

Мощность покрытия может быть в том числе оценена по VC-размерности семейства

Теорема 2.4.4 (Haussler). *Для семейства $\mathcal{F}$, $VCdim(\mathcal{F}) = h$, $\forall \varepsilon \in (0, 1)$*

$$N(\mathcal{F}, \varepsilon, X) \;\leqslant\; Ch(4e)^h \varepsilon^{-h}.$$

Отметим, что оценки основанные на покрытиях, учитывают, до некоторой степени, сходство алгоритмов в семействе, рассматривая алгоритмы ε-шара как один

алгоритм. Однако это делается ценой огрубления порога переобучения. Кроме того, к алгоритмам покрытия, как и в VC-оценке применяется неравенство Буля.

2.5 Вещественно-значные семейства и fat-размерность

Оценки основанные на покрытии работают и в случае вещественно-значных семейств $\mathcal{F}$, когда мощность проекции $\mathcal{F}_{\mathbb{X}}$ может быть бесконечна и VC-оценка, основанная на коэффициенте разбиения $|\mathcal{F}_{\mathbb{X}}|$, становится неприменима [13].

В этом случае на семействе $\mathcal{F}$ определяется метрика вида

$$\rho_1(f, f') = \sum_{x \in X} |f(x) - f'(x)| .$$

Аналогично случаю бинарной классификации, определяется минимальная мощность покрытия $N_1(\mathcal{F}, \varepsilon, X)$ и рассматривается ее максимум по всевозможным обучающим выборкам:

$$N_1(\mathcal{F}, \varepsilon, \ell) = \max \{N_1(\mathcal{F}, \varepsilon, X) \colon X \subset \mathcal{X}\}$$

Аналогично тому, как в случае бинарной классификации мощность покрытия оценивалась сверху через VC-размерность, в данном случае она оценивается при помощи fat-размерности (fat-dimension) введенной в теорию обучения в работе [40].

Напомним, что вещественнозначное семейство $\mathcal{F}$ разбивает (shatter) выборку X, если существуют такие пороги $r_1, \ldots, r_\ell$, что для любого $b \in \{-1, 1\}^\ell$ найдется $f \in \mathcal{F}$ такая, что $b_i f(x_i) > b_i r_i$. Понятие ε-разбиения обобщает это определение добавляя «зазор» ε.

Определение 2.5.1. *Вещественнозначное семейство $\mathcal{F}$ ε-разбивает (ε-shatter) выборку X, если существуют такие пороги $r_1, \ldots, r_\ell$, что для любого $b \in \{-1, 1\}^\ell$ найдется $f \in \mathcal{F}$ такая, что $b_i f(x_i) > b_i r_i + \varepsilon$.*

Определение 2.5.2. *fat-размерность $fat_{\mathcal{F}}(\varepsilon)$ семейства $\mathcal{F}$ при заданном ε есть максимальная длина выборки X, которая ε-разбивается семейством.*

Fat-размерность относят к классу размерностей зависящих от масштаба (scale sensitive dimension), несколько видов таких размерностей исследуются в [59].

Теорема 2.5.3 ([59]). *Для вещественнозначного семейства $\mathcal{F}$ с областью значений $[0, 1]$, $\varepsilon \in (0, 1]$ и fat-размерностью $fat_{\mathcal{F}}(\varepsilon/8) = d$ при всех $\ell \geqslant d$ верно*

$$N_1(\mathcal{F}, \varepsilon, \ell) \;<\; 2 \left(\frac{4}{\varepsilon}\right)^{3d \log_2(16e\ell/(d\varepsilon))} .$$

Оценка обобщающей способности для вещественнозначных семейств похожа на оценку для бинарных классификаторов.

Теорема 2.5.4. *Если $\mathcal{F}$ семейство вещественнозначных функций $\mathcal{X} \to [0,1]$ и $\mathcal{D}$ неизвестное распределение на $\mathcal{X} \times [0,1]$ и $\varepsilon \in (0,1)$, то*

$$\mathbf{P}\Big[\exists f \in \mathcal{F}\colon \Big|L_f - \hat{L}_f\Big| \geqslant \varepsilon\Big] \leqslant 4N_1(\varepsilon/16, \mathcal{F}, 2\ell)e^{-\varepsilon^2\ell/32}$$

2.6 Рademахеровская сложность

Еще одна мера сложности семейства $\mathcal{F}$ помимо рассмотренных выше мощности покрытия, VC- и fat-размерности — Рademахеровская сложность [45],[44], [46].

Рademахеровской случайной величиной называется $\sigma \in \{-1,+1\}, \mathbf{P}[\sigma = -1] = \mathbf{P}[\sigma = +1] = 0.5$, будем обозначать $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_\ell)$ вектор таких величин. Рademахеровская сумма для $f \in \mathcal{F}$ и $X \in \mathcal{X}^\ell$ есть выборочная ковариация значений функции на объектах выборки X с реализацией вектора рademахероских величин $\boldsymbol{\sigma}$:

$$\mathcal{R}(f, X, \boldsymbol{\sigma}) = \left|\sum_{i=1}^{\ell} \sigma_i f(x_i)\right|.$$

Максимум этой суммы по всем функциям $\mathcal{F}$ показывает, насколько точно можно аппроксимировать функциями из $\mathcal{F}$ заданную реализацию вектора «рademахероского шума» $\boldsymbol{\sigma}$:

$$\mathcal{R}(\mathcal{F}, X, \boldsymbol{\sigma}) = \sup_{f \in \mathcal{F}} \mathcal{R}(f, X, \boldsymbol{\sigma}).$$

Если для любого значения вектора шума $\boldsymbol{\sigma}$ можно найти в $\mathcal{F}$ функцию достаточно точно его аппроксимирующую, то среднее по $\boldsymbol{\sigma}$ — эмпирическая или выборочная рademахеровская сложность семейства — будет велика:

$$\mathcal{R}(\mathcal{F}, X) = \mathbf{E}_{\boldsymbol{\sigma}}\mathcal{R}(\mathcal{F}, X, \boldsymbol{\sigma}) = \frac{1}{2^\ell} \sum_{\boldsymbol{\sigma} \in \{-1,+1\}^\ell} \mathcal{R}(\mathcal{F}, X, \boldsymbol{\sigma}).$$

Рademахеровская сложность семейства есть

$$\mathcal{R}(\mathcal{F}) = \mathbf{E}\mathcal{R}(\mathcal{F}, X).$$

Последняя величина может быть оценена сверху при помощи VC- и fat-размерностей [53].

Рademахеровская сложность используется в оценках аналогично VC- или fat-размерности. Существенным улучшением здесь является то, что для получения оценки равномерной сходимости не используется неравенство Буля — один из основных факторов завышенности других оценок.

Теорема 2.6.1 ([18]). *Пусть функция потерь L принимает значения в интервале $[0,1]$ и $\mathcal{L} = \{L_f \colon f \in \mathcal{F}\}$ — семейство функций потерь индуцированное $\mathcal{F}$. Тогда*

$$\mathbf{P}\left[\sup_{f\in\mathcal{F}}\left|L_f - \hat{L}_f\right| > 2\mathcal{R}(\mathcal{L}) + \sqrt{2/\ell \ln(1/\eta)}\right] \leqslant \eta$$

и

$$\mathbf{P}\left[er(f) - \hat{er}(f) > 2\mathcal{R}(\mathcal{L}, X) + \sqrt{2/\ell \ln(2/\eta)}\right] \leqslant \eta.$$

Отметим, что вторая оценка может быть вычислена (хотя это и достаточно трудоемко) на основе наблюдаемой обучающей выборки. Оценивание радемахеровской сложности $\mathcal{R}(\mathcal{L})$ или $\bar{\mathcal{R}}(\mathcal{L}) = \sup\limits_{\mathcal{D}} \mathcal{R}(\mathcal{L})$ представляет отдельную задачу, решаемую для конкретных семейств аналогично задаче определения VC-размерности.

Рассматривают также локальную радемахеровскую сложность семейства. Если определить

$$\mathcal{R}(\mathcal{F}, X, \boldsymbol{\sigma}, r) = \sup_{f\colon \mathbf{E}f^2 \leqslant r} \mathcal{R}(f, X, \boldsymbol{\sigma}, r),$$

то локальная сложность $\mathcal{R}(\mathcal{F}, r) = \mathbf{E}_{\mathcal{D},\boldsymbol{\sigma}}\mathcal{R}(\mathcal{F}, X, \boldsymbol{\sigma}, r)$ есть сложность части семейства $\mathcal{F}$ состоящей из функций с небольшой вариацией, среди которых меньше проявляется эффект переобучения. Пример оценки использующей локальную сложность можно найти в [19].

2.7 Оценки с использованием понятия отступа

Для задачи бинарной классификации $Y = \{-1, +1\}$. естественно использовать семейство бинарных функций $\mathcal{F}\colon \mathcal{X} \to Y$ и число ошибок функции $\sum_i [\, f(x_i) \neq y_i]$ в качестве критерия выбора функции $\hat{f}$ в методе обучения.

На практике многие методы обучения используют для бинарной классификации вещественнозначные семейства $\mathcal{F}$ и некоторую непрерывную функцию потерь $\hat{L}_f = 1/\ell \sum_i L(f(x_i), y_i)$ в качестве критерия выбора $\hat{f} \in \mathcal{F}$. Классификация производится по пороговому правилу вида $\operatorname{sign}(\hat{f}(x))$.

Функция потерь L на объекте x_i может зависеть от *отступа*(margin) функции f на x_i:

Определение 2.7.1. *Отступом функции $f \in \mathcal{F} : \mathcal{X} \to Y$ называется величина $M_f(x_i, y_i) = y_i f(x_i)$. Отступ положителен, если $\operatorname{sign}(f(x_i))$ правильно классифицирует прецедент (x_i, y_i) и отрицателен иначе.*

Отступ можно интерпретировать как расстояние объекта x_i до разделяющей поверхности $f(x) = 0$ и как показатель «надежности» или «степени уверенности» классификации x_i функцией f. Функция потерь может быть определена, к примеру, как $L(f(x_i), y_i) = (1 - f(x_i)y_i)_+$, как в методе опорных векторов (SVM) или $L(f(x_i), y_i) = = (1 - f(x_i)y_i)^2$, как в методе линейного дискриминанта Фишера. Средний отступ обучающей выборки определяется как $\hat{L}_f = 1/\ell \sum_i L(f(x_i), y_i)$.

Интуитивно представляется, что функция f с бóльшим средним отступом на обучающей выборке должна обладать лучшей обобщающей способностью и делать меньше ошибок на контрольной выборке из $\mathcal{D}$, чем функция f' с тем же числом ошибок на обучении что и f, но маленьким средним отступом. Минимизация функции потерь зависящей от отступа может объяснять приемлемое качество обучения при использовании семейств высокой сложности (с бесконечной VC-размерностью) или большим числом параметров (больше размера обучающей выборки). Оценки обобщающей способности основанные на понятии отступа формализуют эти интуитивные представления.

Для вещественнозначного семейства $\mathcal{F}$ естественной мерой сложности является мощность ε-покрытия рассмотренная выше. Пусть расстояние между функциями определяется как

$$\rho_\infty(f, f', X) = \max_i |f(x_i) - f'(x_i)| .$$

Пусть минимальная мощность покрытия $N_\infty(\mathcal{F}, \varepsilon, X)$ и ее максимум по обучающим выборкам $N_\infty(\mathcal{F}, \varepsilon, \ell)$ определены аналогично тому, как это сделано в предшествующем параграфе. Пусть $\nu(f) = \mathbf{P}[\, yf(x) < 0]$ есть вероятность ошибки f и

$$\hat{\nu}_\gamma(f) = \frac{1}{\ell} \sum_i [\, y_i f(x_i) < \gamma] , \ \gamma > 0$$

есть доля объектов в обучающей выборке для которых отступ f менее γ. Последнее есть более строгое частоты ошибки, чем стандартное $\hat{\nu}(f) = \frac{1}{\ell} \sum_i [\, y_i f(x_i) < 0]$ — очевидно $\hat{\nu}(f) \leqslant \hat{\nu}_\gamma(f)$.

Тогда выполняется следующая оценка.

Теорема 2.7.2 ([13]). *Пусть $\mathcal{F}\colon \mathcal{X} \to \mathbb{R}$ семейство вещественнозначных функций, тогда для любого (фиксированного) $\gamma > 0$ и $\varepsilon \in (0, 1)$ верна оценка:*

$$\mathbf{P}\Big[\exists f \in \mathcal{F}\colon \nu(f) \geqslant \hat{\nu}_\gamma(f) + \varepsilon \Big] \ \leqslant \ 2N_\infty(\mathcal{F}, \gamma/2, 2\ell)\, e^{-\ell\varepsilon^2/8}$$

Отметим, что в этой оценке отступ γ который мы выбираем, задает масштаб покрытия определяющий сложность семейства. Чем больше отступ, тем больше масштаб и тем меньше мощность покрытия. Этот факт может быть интерпретирован

таким образом, что если метод μ выбирает из функций с большим отступом на обучающей выборке, то, эффективно, это делает сложность используемого им семейства меньше и уменьшает оценку вероятности переобучения.

2.8 PAC-Bayes подход

PAC-подход и байесовский подход к обучению нередко противопоставляются друг другу. Методы обучения, рассматриваемые в PAC-подходе, обычно дают на выходе функцию $\hat{f}$ классификации/регрессии из некоторого семейства $\mathcal{F}$ и нас интересуют *доверительные интервалы* для вероятности ошибки L_f, не зависящие от распределения генерирующего обучающие данные. Чтобы исключить зависимость доверительной оценки от метода обучения, обычно рассматривают максимальный доверительный интервал по всему семейству функций.

Методы обучения в байесовском подходе дают на выходе *распределение* Q на семействе $\mathcal{F}$ зависящее от априорного распределения P на $\mathcal{F}$, данных X и предположения о вероятностной модели, генерирующей данные. Ответ для объекта $x \in \mathcal{X}$ при этом обычно получается как среднее ответов функций из $\mathcal{F}$ с усреднением по апостериорной мере Q. Естественной мерой качества при этом является *средняя* вероятность ошибки $\mathbf{E}_Q L(f)$ по апостериорной мере на семействе. Проблема переобучения, то есть в данном случае зависимость качества ответов от выбора P и вероятностой модели данных, обычно не рассматривается.

PAC-Bayes подход является плодотворной попыткой применения PAC-подхода к получению оценок обобщающей способности для байесовских методов обучения. Основная PAC-Bayes оценка при этом оказывается верна для любого апостериорного распределения Q, не обязательно полученого по формуле Байеса из P, X (но возможно зависящего от них).

Пусть $Y = \{-1, +1\}$ и $L = [\, f(x) \neq y]$.

Теорема 2.8.1. *Для любого априорного распределения P на $\mathcal{F}$ и $\eta \in (0, 1)$*

$$\mathbf{P}_X \left[\forall Q,\, \mathbf{E}_Q \nu(f) \geqslant \mathbf{E}_Q \hat{\nu}(f) + \sqrt{\frac{KL(Q||P) + \ln(\ell/\eta) + 2}{2\ell - 1}} \right] \leqslant \eta,$$

где $KL(Q||P) = \mathbf{E}_Q \ln \frac{Q(f)}{P(f)}$ — дивергенция Кульбака-Лейблера между распределениями Q и P.

Если выбираемое методом μ апостериорное распределение близкó к априорному, то есть метод не настраивается на обучающие данные, то дивергенция $KL(Q||P)$ мала, что делает второе слагаемое правой части меньше и оценку точнее. С другой

стороны, при этом не минимизируется первое слагаемое оценки — средняя ошибка на обучающей выборке. Настраиваясь на обучающие данные, метод уменьшает первое слагаемое, при этом, возможно, увеличивая второе. Таким образом, в процессе обучения мы получаем возможность распорядится свободой выбора Q для минимизации величины оценки для средней ошибки $\mathbf{E}_Q\nu(f)$. На практике необходимо задаться некоторым параметрическим видом для P, Q, чтобы такая минимизация была вычислительно эффективна. При этом возможно получение достаточно точных оценок и эффективных методов обучения основанных на них, известный пример такого рода — PAC-Bayes оценка для линейных классификаторов [50].

Глава 3

Оценки на основе характеристик расслоения семейства

В настоящей главе рассматриваются shell-оценки обобщающей способности, основанные на следующем соображении. Обычно бóльшая часть алгоритмов в $\mathcal{F}$ имеет вероятность ошибки (или частоту ошибок на полной выборке) около 50%. Если метод обучения выбирает алгоритм с малой частотой ошибок на обучающей выборке, то фактически выбор производится не из всего семейства $\mathcal{F}$, а лишь из очень небольшой его части, состоящей из алгоритмов с малой вероятностью ошибки. Размер этой части семейства существенно ниже размера всего семейства, что предполагает возможность точнее оценить вероятность переобучения по сравнению, скажем, с классическими оценками Вапника-Червоненкиса.

В настоящей главе в более простом и общем виде выводятся комбинаторные аналоги shell-оценок из [49, 47], и проводится анализ причин их завышенности. Показывается их связь с VC-оценкой и оценкой «бритвы Оккама».

В параграфе 3.3 выводится комбинаторная shell-оценка, которая учитывает эффект расслоения семейства, но требует знания полной выборки, то есть является ненаблюдаемой. В параграфе 3.4 выводится наблюдаемая комбинаторная shell-оценка, основанная на расслоении семейства по частоте ошибок на обучающей выборке.

3.1 Профили расслоения семейства алгоритмов

Идея shell-оценок [47] состоит в том, чтобы учесть распределение истинной частоты ошибок алгоритмов в семействе $\mathcal{F}$. В комбинаторном подходе мы будем, соответственно, рассматривать *профиль истинных частот* ошибок множества A (или

профиль расслоения [6]):

$$\Delta_m = \operatorname{card}\{a \in A\colon n_a = m\}\,.$$

Будем называть соответствующее подмножество $A_m = \{a \in A\colon n_a = m\}$ m-ым *слоем* множества A. Профиль Δ_m есть совместная характеристика семейства $\mathcal{F}$ и полной выборки $\mathbb{X}$. Определим также *профиль наблюдаемых частот* ошибок на обучающей выборке X:

$$\hat{\Delta}_s = \operatorname{card}\{a \in A\colon \hat{n}_a = s\}\,.$$

Отметим, что профиль $\hat{\Delta}_s$, в отличие от Δ_m, есть случайная величина. На Рис. 3.1 приведены примеры оценки профилей для семейства $\mathcal{F}$ линейных классификаторов по методу Монте-Карло на некоторой случайно сгенерированной выборке $\mathbb{X}$. Отметим, что симметричность профилей относительно частоты $\frac{1}{2}$ обусловлена свойствами данного семейства $\mathcal{F}$ (линейных классификаторов) и не зависит от полной выборки. Более точно, профили симметричны для любого семейства, в котором для любого алгоритма найдется его «инверсия» — алгоритм с инвертированным вектором ошибок.

Профиль $\hat{\Delta}_s$, рассчитанный для случайной выборки X, по форме практически повторяет Δ_m; более высокий абсолютный уровень $\hat{\Delta}_s$ и его большая «гладкость» обусловлены меньшим числом $(\ell + 1)$ возможных значений наблюдаемой ошибки в сравнении с $L + 1$ возможными значениями полного числа ошибок.

Отметим также, что $\mathbf{E}\hat{\Delta}_s = \sum_{m=0}^{L} \Delta_m h_m(s)$, то есть $\hat{\Delta}_s$ не является несмещенной оценкой профиля расслоения Δ_m в точке $m = \frac{L}{\ell}s$, но представляет некоторое «усреднение» этого профиля в окрестности $m = \frac{L}{\ell}s$, пика гипергеометрического распределения $h_m(s)$.

3.2 Обзор работ по теме

Одна из первых попыток учета расслоения семейства по истинной частоте ошибок в оценках обобщающей способности была предпринята в [58], где предлагалось отказаться от огрубления $\max\limits_{m}$ в оценке Вапника-Червоненкиса:

$$\mathbf{P}[\,\nu_{\hat{a}} - \hat{\nu}_{\hat{a}} \geqslant \varepsilon] \;\leqslant\; \sum_{a \in A} H_{n_a}(s_{n_a}) \;=\; \sum_{m=0}^{L} \Delta_m H_m(s_m)\,;$$

Поведение последней величины изучается в [58] с использованием методов статистической физики. В этой же работе вводится термин «расслоение» (shell decomposition

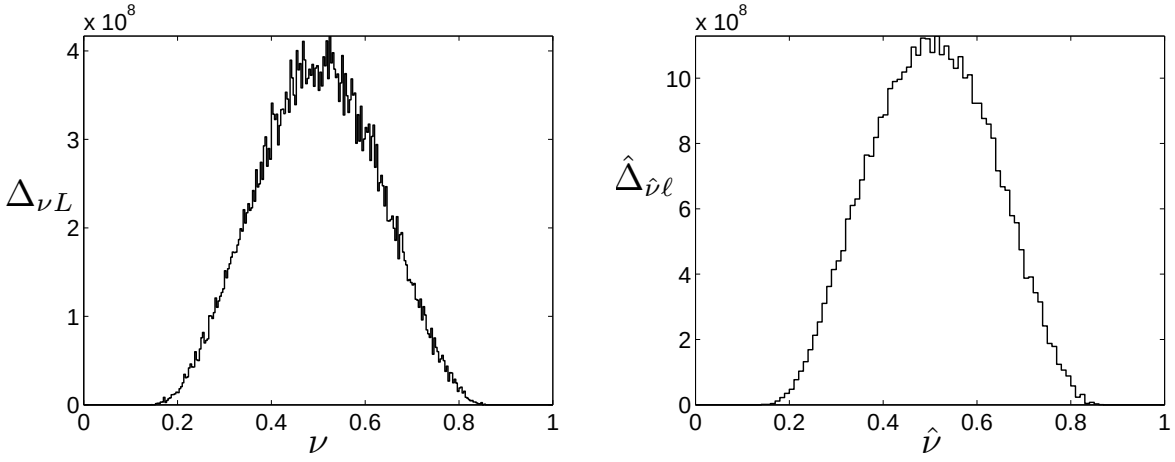

Рис. 3.1. Монте-Карло оценки профиля истинных частот $\Delta_{\nu L}$ и профиля наблюдаемых частот $\hat{\Delta}_{\hat{\nu}\ell}$ (последний — на случайно взятой обучающей выборке $X \in [\mathbb{X}]^\ell$). Для семейства $\mathcal{F}$ линейных классификаторов в $\mathbb{R}^5$ и случайной двухклассовой полной выборки $\mathbb{X} \subset \mathbb{R}^5$, взятой из смеси нормальных распределений; $L = 300$, $\ell = 100$. Степень перекрытия классов в $\mathbb{X}$ (вероятность ошибки байесовского классификатора) $\approx 15\%$. Полный размер множества алгоритмов $|A| \approx 4 \cdot 10^{10}$, размер выборки алгоритмов в методе Монте-Карло $\tilde{A} \subset A$, $|\tilde{A}| = 10^5$.

of the union bound), под которым понимается группировка слагаемых в неравенстве Буля по значениям полного числа ошибок $n_a = m$.

В [60] рассматривается метод обучения МЭР (минимизация эмпирического риска) $\mu(X) = \arg\min_a \hat{n}_a$, и в предположении, что профиль Δ_m известен, оценивается среднее значение истинной частоты ошибок алгоритма, получающегося в результате обучения: $\mathbf{E}n_{\hat{a}}$. Для получения такой оценки $\mathbf{E}n_{\hat{a}}$, которая зависела бы только от наблюдаемой выборки, профиль Δ_m предлагается заменять в оценке на профиль проекции $I(\mathcal{F}, X)$ семейства $\mathcal{F}$ на обучающую выборку X. Полученная для $\mathbf{E}n_{\hat{a}}$ оценка далее используется для выбора оптимального семейства $\mathcal{F}$. К сожалению, для получения оценки делается нереалистичное предположение, что все случайные величины $\{\hat{n}_{\hat{a}} \colon a \in A\}$ независимы. На практике имеет место обратная ситуация — в A много схожих (по метрике Хэмминга) алгоритмов, и $\hat{n}_{\hat{a}}$ существенно зависимы.

В [49] и [47] выводятся доверительные оценки для $n_{\hat{a}}$, использующие Δ_m и $\hat{\Delta}_s$. Далее мы получим аналоги этих оценок в рамках комбинаторного подхода и покажем их связь с оценкой Вапника-Червоненкиса и оценкой «бритвы Оккама».

3.3 Shell-оценки зависящие от полной выборки

Следующая теорема показывает, что shell-оценка [49] есть частный случай оценки «бритвы Оккама», если взять одинаковые пороги переобучения η_a для алгоритмов с одинаковым полным числом ошибок: $\eta_a = \eta(n_a)$. В этом случае мы оцениваем функционал вероятности переобучения

$$Q_{shell1} \stackrel{\text{def}}{=} \mathbf{P}[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta(n_{\hat{a}})]\,,$$

зависящий от L порогов переобучения $\boldsymbol{\eta} = \Big(\eta(m)\colon m = 1, \ldots, L\Big)$.

Теорема 3.3.1. *Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$ длины L, метода обучения μ, индикатора ошибки I, если профиль расслоения множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ есть Δ_m, то выполняется оценка вероятности переобучения:*

$$\mathbf{P}[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \;\leqslant\; \eta(n_{\hat{a}})] \;\leqslant\; \sum_{m=0}^{L} \Delta_m \eta(m).$$

При выборе порогов переобучения вида $\eta(m) = \eta(m, \alpha) = \frac{\alpha}{L\,\Delta_m}$ также имеет место доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}_1(\hat{n}_{\hat{a}}, \boldsymbol{\eta}(\alpha))] \;\leqslant\; \alpha,$$

где $\bar{n}_1(\hat{n}_{\hat{a}}, \boldsymbol{\eta}(\alpha)) = \min\{m : H_m(\hat{n}_{\hat{a}}) \leqslant \eta(m, \alpha)\}$ и при условии, что множество в определении $\bar{n}_1$ имеет форму интервала.

Доказательство. Из оценки «бритвы Оккама» (1.15) и определения профиля Δ_m следует

$$\mathbf{P}[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \;\leqslant\; \eta(n_{\hat{a}})] \;\leqslant\; \sum_{a \in A} \eta(n_a) \;=\; \sum_{m=0}^{L} \Delta_m \eta(m).$$

Определяя пороги переобучения так, чтобы правая часть была равна α, к примеру,

$$\boldsymbol{\eta}(\alpha) = \left(\eta(m, \alpha) = \tfrac{\alpha}{L\,\Delta_m}\colon m = 1, \ldots, L\right),$$

имеем $\mathbf{P}[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta(n_{\hat{a}}, \alpha)] \leqslant \alpha$.

Для получения доверительной оценки $n_{\hat{a}}$, решим неравенство под знаком вероятности относительно $n_{\hat{a}}$. Заметим, что в отличие от (1.10), здесь $n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta(n_{\hat{a}}, \alpha))$ не является его решением, поскольку η зависит от величины, по которой идет минимизация в $\bar{n}\,(\hat{n}_{\hat{a}}, \eta)$. Определим функцию

$$\bar{n}_1(\hat{n}, \boldsymbol{\eta}) = \min\{m : H_m(\hat{n}) \leqslant \eta(m)\}\,,$$

тогда

$$H_n(\hat{n}) \leqslant \eta(n,\alpha) \quad \Leftrightarrow \quad n > \bar{n}_1(\hat{n}, \boldsymbol{\eta}(\alpha)),$$

а значит $\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}_1(\hat{n}_{\hat{a}}, \boldsymbol{\eta}(\alpha))] \leqslant \alpha$, получаем утверждение теоремы. ■

В теореме предполагается, что неравенство $H_m(s) \leqslant \frac{\alpha}{L\Delta_m}$ относительно m имеет решение в форме интервала. То есть, в произведении $H_m(s)\,\Delta_m$ функция $H_m(s)$ падает с ростом m достаточно быстро, чтобы компенсировать рост профиля Δ_m. Известно, что $H_m(s)$ с увеличением m при $m > sL/\ell$ падает экспоненциально быстро, профиль же, как показывает, в частности, эксперимент с линейными классификаторами, растет более умеренно. Если это не так для каких-то семейств $\mathcal{F}$, то могут возникать «фазовые переходы» [49] — разрывы в кривой зависимости оценки $\bar{n}_1(\hat{n}_{\hat{a}}, \boldsymbol{\eta}(\alpha))$ от α.

Мотивацией для вывода комбинаторных аналогов shell-оценок послужило, в частности, следующее утверждение [47]: «Shell оценка *существенно* более точна, чем оценка "бритвы Оккама" за счет того, что она учитывает *существенно* больше информации о задаче», то есть профиль Δ_m. Данное утверждение представляется неверным. В действительности последняя оценка является частным случаем оценки «бритвы Оккама» с дополнительным ограничением на выбор порогов $\eta_a = \eta(n_a)$; поэтому лучшее значение оценки, которого можно добиться выбором порогов для неё, заведомо хуже, чем у оценки «бритвы Оккама».

Далее, shell-оценка из [47, глава 8] является вариантом оценки Вапника-Червоненкиса, в которой порогу η разрешено зависеть от $\hat{n}_{\hat{a}}$. То есть рассматривается функционал вероятности переобучения

$$Q_{shell2} \stackrel{\text{def}}{=} \mathbf{P}\Big[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta(\hat{n}_{\hat{a}})\Big]\,,$$

зависящий от $\ell+1$ порогов $\boldsymbol{\eta} = (\eta(s)\colon s = 0, \ldots, \ell)$.

Для получения ее комбинаторного аналога определим функцию

$$P(s,\eta) \stackrel{\text{def}}{=} \sum_{m \geqslant \bar{n}\,(s,\eta)} \Delta_m h_m(s)\,,$$

которая является верхней оценкой вероятности того, что хотя бы один алгоритм из A имеет на обучающей выборке s ошибок и при этом переобучен:

$$\mathbf{P}\Big[\, \exists a \in A\colon \hat{n}_a = s \;\wedge\; n_a \geqslant \bar{n}\,(s,\eta)\Big] \;\leqslant\; P(s,\eta).$$

Утверждение 3.3.2. *Функция $P(s,\eta)$ не убывает по η.*

Теорема 3.3.3. *Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, $|\mathbb{X}| = L$, метода обучения μ, индикатора ошибки I, если профиль расслоения множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ есть Δ_m, то имеет место оценка вероятности переобучения*

$$\mathbf{P}\Big[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta(\hat{n}_{\hat{a}})\Big] \;\leqslant\; \sum_{s=0}^{\ell} P(s, \eta(s)).$$

Верна также доверительная оценка:

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta(\hat{n}_{\hat{a}}, \alpha))] \;\leqslant\; \alpha,$$

где $\eta(s, \alpha) = \max\big\{\eta\colon P(s, \eta) \leqslant \frac{\alpha}{\ell}\big\}$.

Доказательство. Аналогично VC-оценке (Теорема 1.5.1),

$$\mathbf{P}[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta(\hat{n}_{\hat{a}})] \;\leqslant\; \mathbf{P}\Big[\, \exists a \in A\colon H_{n_a}(\hat{n}_a) \;\leqslant\; \eta(\hat{n}_a)\Big] \leqslant$$
$$\leqslant \sum_{a \in A} \mathbf{P}[\, H_{n_a}(\hat{n}_a) \leqslant \eta(\hat{n}_a)] \;=\; (P).$$

Далее, очевидно,

$$\mathbf{P}[\, H_{n_a}(\hat{n}_a) \leqslant \eta(\hat{n}_a)] \;=\; \sum_{s=0}^{\ell} [\, H_{n_a}(s) \leqslant \eta(s)]\, \mathbf{P}[\, \hat{n}_a = s]\,;$$

и по определению: $\mathbf{P}[\, \hat{n}_a = s] \;=\; h_{n_a}(s)$. Следовательно, имеем:

$$(P) = \sum_{a \in A} \sum_{s} [\, H_{n_a}(s) \leqslant \eta(s)]\, h_{n_a}(s) =$$
$$= \sum_{s=0}^{\ell} \sum_{m=0}^{L} \Delta_m\, [\, H_m(s) \leqslant \eta(s)]\, h_m(s) = \sum_{s=0}^{\ell} P(s, \eta(s)).$$

Здесь мы воспользовались определением профиля Δ_m и функции $P(s, \eta)$. Решая неравенство под знаком вероятности в $\mathbf{P}[\, H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \eta(\hat{n}_{\hat{a}})]$ относительно $n_{\hat{a}}$, имеем:

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta(\hat{n}_{\hat{a}}))] \;\leqslant\; \sum_{s=0}^{\ell} P(s, \eta(s)).$$

Фиксируем некоторую доверительную вероятность α и определим вектор порогов $\boldsymbol{\eta}$ так, чтобы правая часть была равна α. К примеру,

$$\eta(s, \alpha) \;\stackrel{\text{def}}{=}\; \max\big\{\eta\colon P(s, \eta) \leqslant \tfrac{\alpha}{\ell}\big\}\,.$$

Имеем утверждение теоремы. ■

Легко видеть, что если в последней теореме взять одинаковые пороги $\eta(s) = \eta_0$, то мы получим VC-оценку:

$$\sum_{s=0}^{\ell} P(s, \eta_0) = |A|\, \eta_0 = \alpha_0.$$

Если при этой же доверительной вероятности α_0 выбрать пороги $\eta(s, \alpha_0)$ так, как это сделано в теореме, то часть из них будет больше η_0, а часть меньше. Это означает, что для некоторых наблюдаемых частот теорема дает лучшую оценку, чем оценка Вапника-Червоненкиса, для других — худшую.

Как и оценка «бритвы Оккама», эта оценка не устраняет основных факторов завышенности VC-оценки, но может быть точнее последней *в среднем* при правильном подборе вектора порогов $\Big(\eta(s)\Big)$ — таких, чтобы оценка $\bar{n}\,(\hat{n}_{\hat{a}}, \eta(\hat{n}_{\hat{a}}))$ была меньше для значений $\hat{n}_{\hat{a}} = s$, которые чаще получаются в результате обучения.

Определение 3.3.4. *Распределение частоты ошибок на обучении, индуцируемое методом обучения μ есть*

$$P(\hat{\nu}) \stackrel{\text{def}}{=} \mathbf{P}[\,\hat{\nu}_{\hat{a}} = \hat{\nu}] \;=\; \mathbf{P}[\,\nu(\mu X, X) = \hat{\nu}]\,,$$

где $\hat{\nu} \in \left\{\frac{0}{\ell}, \ldots, \frac{\ell}{\ell}\right\}$.

На Рис. 3.2 приведен пример оценки распределения $P(\hat{\nu})$ по методу Монте-Карло; отметим, что частота ошибок на обучении сконцентрирована в области малых частот.

Если профиль Δ_m сконцентрирован возле частоты $\nu_a = \frac{1}{2}$, то функции $P(s, \eta)$ меньше для малых значений s, чем для средних и, следовательно, последняя оценка должна быть более точна для малых s, которые чаще получаются при обучении.

Определение оптимальных порогов аналогично Лемме 1.6.1 в данном случае более проблематично, поскольку мы имеем задачу минимизации

$$\min_{(\eta(s))} \sum_{s=0}^{\ell} p(s)(-\ln \eta(s)), \quad \text{при условии} \quad \sum_{s=0}^{\ell} P(s, \eta(s)) \leqslant \alpha,$$

с нелинейными дискретными ограничениями (вместо $\sum\limits_{a \in A} \eta_a = \alpha$ в Лемме 1.6.1).

Рассмотрим простейший вариант выбора порогов, в некотором смысле наилучший для получения точной оценки для малых частот на обучении. Предположим, что нам известно, что частота ошибок на обучении у алгоритма $\hat{a}$ не выше некоторого порога: $\mathbf{P}[\,\hat{n}_{\hat{a}} > s_0] = 0$. В примере на Рис. 3.2 это $\approx 20\%$. Тогда мы можем положить в последней теореме $\forall s > s_0 \colon \eta(s) = 0$, то есть пренебречь оценкой для частот выше s_0/ℓ, но за счет этого получить более точную оценку для малых частот при сохранении прежнего уровня значимости α.

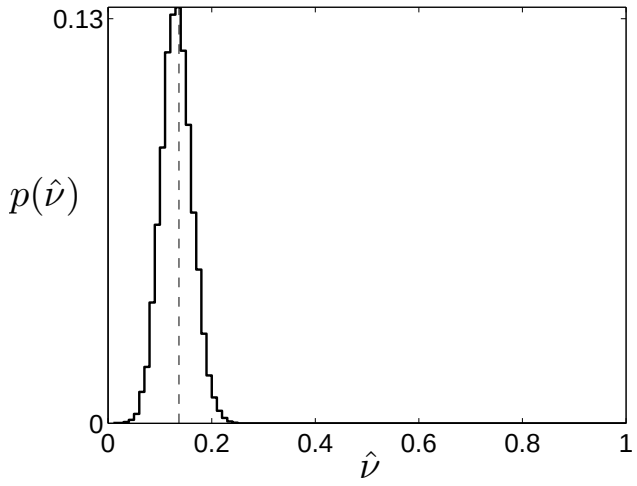

Рис. 3.2. Монте-Карло оценка распределения $P(\hat{\nu})$, индуцированного методом μ – машиной опорных векторов (SVM). Использованы та же полная выборка $\mathbb{X}$ и семейство $\mathcal{F}$, что в оценках профилей на Рис. 3.1. Обучение методом SVM произведено на 10 000 случайных обучающих выборок $X \in [\mathbb{X}]^\ell$. На рисунке представлена гистограмма частот алгоритмов, получившихся в результате обучения; вертикальный пунктир отмечает частоту ошибок алгоритма, получаемого при обучении на полной выборке $\mathbb{X}$.

Следствие 3. *В условиях предыдущей теоремы справедлива оценка*

$$\mathbf{P}[\, n_{\hat{a}} < \bar{n}\,(\hat{n}_{\hat{a}}, \eta(\hat{n}_{\hat{a}}, \alpha))] \;\geqslant\; 1 - \alpha,$$

где $\eta(s,\alpha) = \max\left\{\eta\colon P(s,\eta) \leqslant \frac{\alpha}{s_0}\right\}$ *при* $s \leqslant s_0$, *иначе* $\eta(s,\alpha) = 0$.

3.4 Shell-оценка, зависящая от обучающей выборки

В оценке Теоремы 3.3.3 функции $P(s,\eta)$ зависят от полного числа ошибок алгоритмов $a \in A$. Получим для $P(s,\eta)$ оценку $\hat{P}(s,\eta)$, зависящую от числа ошибок алгоритмов на обучающей выборке X. Нам необходима верхняя оценка, поскольку тогда, заменяя в Теореме 3.3.3 P на $\hat{P}$, мы получим более консервативные пороги переобучения

$$\hat{\eta}(s,\alpha) = \max\left\{\eta\colon \hat{P}(s,\eta) \leqslant \frac{\alpha}{\ell}\right\},$$

и оценка теоремы будет по прежнему верна, хотя и станет менее точной.

Оценка $\hat{P}(s,\eta)$ зависит от X, то есть является случайной величиной, поэтому мы можем говорить лишь о доверительной верхней оценке для $P(s,\eta)$.

Чтобы получить такую оценку, заметим, что функцию P можно записать в виде

$$P(s,\eta) = \sum_{a\in A} [\, n_a \geqslant \bar{n}\,(s,\eta)]\, h_{n_a}(s)\,.$$

Хотя параметры n_a нам неизвестны, но для каждого из них есть доверительные интервалы

$$[\,\underline{n}\,(\hat{n}_a, \delta/2),\, \bar{n}\,(\hat{n}_a, \delta/2)] \stackrel{\text{def}}{=} C(\hat{n}_a, \delta);$$

и из Лемм 1.4.1 и 1.4.2 имеем

$$\mathbf{P}[\,n_a \notin C(\hat{n}_a, \delta)] \leqslant \delta.$$

Рассмотрим некоторую фиксированную выборку X и предположим, что для неё все параметры $n_a, a \in A$ находятся внутри соответствующих доверительных интервалов и найдем, каким при этом могло бы быть в худшем случае значение $P(s, \eta)$:

$$\hat{P}(s, \eta, \delta) \stackrel{\text{def}}{=} \max_{\{n_a \in C(\hat{n}_a, \delta)\}} P(s, \eta) = \tag{3.1}$$

$$= \max_{\{n_a \in C(\hat{n}_a, \delta)\}} \sum_{a \in A} [\,n_a \geqslant \bar{n}\,(s, \eta)]\, h_{n_a}(s) = \tag{3.2}$$

$$= \sum_{a \in A} \max_{n_a \in C(\hat{n}_a, \delta)} [\,n_a \geqslant \bar{n}\,(s, \eta)]\, h_{n_a}(s) = \tag{3.3}$$

$$= \sum_{a \in A} [\,n^*(\hat{n}_a) \geqslant \bar{n}\,(s, \eta)]\, h_{n^*(\hat{n}_a)}(s)\,. \tag{3.4}$$

Здесь

$$n^*(\hat{n}_a) = \min\{\bar{n}\,(\hat{n}_a, \delta/2),\ \max\{\bar{n}\,(s, \eta),\, \underline{n}\,(\hat{n}_a, \delta/2)\}\}$$

есть решение последней задачи максимизации. Такой вид решения следует из того, что когда s находится в левом η-«хвосте» $h_{n_a}(s)$ (а именно по таким алгоритмам идёт суммирование в $P(s, \eta)$), функция $h_{n_a}(s)$ убывает по n_a.

Отметим, что мы можем также записать (3.1) как

$$\hat{P}(s, \eta, \delta) = \sum_{m=0}^{L} \Delta^*_m h_m(s)\,,$$

где случайная величина

$$\Delta^*_m = \sum_{s=0}^{\ell} \hat{\Delta}_s\,[\,n^*(s) = m]$$

есть пессимистическая оценка профиля Δ_m. Таким образом, $\hat{P}(s, \eta, \delta)$ получается из $P(s, \eta)$ заменой профиля Δ_m на его оценку Δ^*_m.

Далее, покажем, что $\hat{P}(s, \eta, \delta)$ определенная в (3.1) дает нам доверительную оценку для $P(s, \eta)$. Для этого докажем следующую лемму.

Лемма 3.4.1. *Пусть* $F(\theta_1,\dots,\theta_k)=\sum_{i=1}^k f_i(\theta_i)$ *есть некоторая аддитивная функция параметров* $\{\theta_i\}$. *Пусть для каждого параметра имеется доверительный интервал* $\hat{C}_i$: $\mathbf{P}\Big[\,\theta_i\in\hat{C}_i\Big]\geqslant 1-\delta$ *и пусть* $\hat{F}=\max\limits_{\{\theta_i\in\hat{C}_i\}}F(\theta_1,\dots,\theta_k)$. *Тогда для любого* $t\in(0,1)$ *верна доверительная оценка*

$$\mathbf{P}\left[\,F(\theta_1,\dots,\theta_k)\leqslant\frac{\hat{F}}{1-t}\right]\;\geqslant\;1-\frac{\delta}{t}.$$

Доказательство основывается на следующих соображениях. Для отдельного параметра θ_i доля выборок, на которых он накрывается доверительным интервалом $\hat{C}_i$, близка к 1 (превышает $1-\delta$). Следовательно, *в среднем* доля параметров в множестве $\{\theta_i\}$, удовлетворяющих этому условию, также близка к 1. В то же время, когда значение параметра накрывается доверительным интервалом, соответствующие слагаемые оценки $\hat{F}$, по построению, больше соответствующих слагаемых F.

Доказательство. По условию леммы имеем $\forall\theta_i$: $\mathbf{P}\Big[\,\theta_i\notin\hat{C}_i\Big]\leqslant\delta$. Складывая все условия с весами f_i/F, получаем

$$\frac{\sum_{i=1}^k f_i(\theta_i)\mathbf{P}\Big[\,\theta_i\notin\hat{C}_i\Big]}{F(\theta_1,\dots,\theta_k)}\leqslant\delta.$$

Меняя местами суммирование и $\mathbf{P}$, получаем

$$\mathbf{E}\frac{\sum_{i=1}^k f_i(\theta_i)\Big[\,\theta_i\notin\hat{C}_i\Big]}{F(\theta_1,\dots,\theta_k)}\leqslant\delta,$$

то есть в среднем вклад в F слагаемых, для которых параметр θ_i оказался вне доверительного интервала, — мал. Обозначим этот вклад через $\hat{\gamma}\in(0,1)$; последнее выражение есть оценка $\mathbf{E}\hat{\gamma}\leqslant\delta$. Из неравенства Маркова имеем $\mathbf{P}[\,\hat{\gamma}>t]\;\leqslant\;\frac{\mathbf{E}\hat{\gamma}}{t}\;\leqslant\;\frac{\delta}{t}$, следовательно,

$$\mathbf{P}\left[\,\frac{\sum_{i=1}^k f_i(\theta_i)\Big[\,\theta_i\notin\hat{C}_i\Big]}{F(\theta_1,\dots,\theta_k)}\;>\;t\right]\;\leqslant\;\frac{\delta}{t}.$$

Используя тождество $\Big[\,\theta_i\notin\hat{C}_i\Big]+\Big[\,\theta_i\in\hat{C}_i\Big]=1$, получаем:

$$\mathbf{P}\left[\,(1-t)F>\sum_{i=1}^k f_i(\theta_i)\Big[\,\theta_i\in\hat{C}_i\Big]\right]\leqslant\frac{\delta}{t}$$

Заметим теперь, что правая часть неравенства под $\mathbf{P}$ по построению меньше $\hat{F}$, следовательно, имеем

$$\mathbf{P}\Big[\,(1-t)F>\hat{F}\Big]\leqslant\frac{\delta}{t}.$$

Обращая знак оценки, получаем утверждение леммы. ■

Следствие 4. *Взяв в качестве F функцию $P(s,\eta)$, зависящую от набора параметров $\{n_a\colon a\in A\}$, в качестве оценки $\hat{F}$ — функцию $\hat{P}(s,\eta,\delta)$, и положив $t=\frac{1}{2}$, получим:*

$$\forall\,\eta\in(0,1),\ \delta\in(0,1),\ s\colon\quad \mathbf{P}\Big[\,P(s,\eta)\leqslant 2\hat{P}(s,\eta,\delta)\Big]\ \geqslant\ 1-2\delta.$$

Утверждение 3.4.2. *Функция $\hat{P}(s,\eta,\delta)$ не убывает по η и не возрастает по δ.*

Доказательство. Для доказательства первой части утверждения достаточно заметить, что функции $P(s,\eta)$ не убывают по η при любых значениях параметров $\{n_a\}$ следовательно, $\hat{P}(s,\eta,\delta)$, как максимум неубывающих функций, также не убывает по η. Для доказательства второй части утверждения можно заметить, что чем больше δ, тем у́же доверительные интервалы $C(\hat{n}_a,\delta)$, по которым идет максимизация. ■

Используя в Теореме 3.3.3 функцию $\hat{P}(s,\eta,\delta)$ вместо $P(s,\eta)$, получим следующую оценку.

Теорема 3.4.3. *Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, $|\mathbb{X}|=L$, метода обучения μ, индикатора ошибки I, справедлива доверительная оценка*

$$\mathbf{P}\Big[\,n_{\hat{a}}<\bar{n}\,(\hat{n}_{\hat{a}},\hat{\eta}(\hat{n}_{\hat{a}},\alpha))\Big]\ \geqslant\ 1-\alpha,$$

где $\hat{\eta}(s,\alpha)=\max\Big\{\eta\colon 2\hat{P}(s,\eta,\frac{\delta}{4\ell})\leqslant\frac{\alpha}{2\ell}\Big\}$.

Доказательство. Определим

$$\eta(s,\alpha)=\max\Big\{\eta\colon P(s,\eta)\leqslant\frac{\alpha}{2\ell}\Big\}$$

— порог переобучения из Теоремы 3.3.3, для того же уровня значимости $\frac{\alpha}{2\ell}$, что в $\hat{\eta}(s,\alpha)$. Отметим, что при фиксированном s порог $\hat{\eta}(s,\alpha)$ есть случайная величина, в отличие от $\eta(s,\alpha)$.

Выборки из $[\mathbb{X}]^\ell$ можно разделить на два типа — для которых $\hat{\eta}(\hat{n}_{\hat{a}},\alpha)\leqslant\eta(\hat{n}_{\hat{a}},\alpha)$ и для которых $\hat{\eta}(\hat{n}_{\hat{a}},\alpha)>\eta(\hat{n}_{\hat{a}},\alpha)$.

Для выборок первого типа

$$\mathbf{P}\Big[\,H_{n_{\hat{a}}}(\hat{n}_{\hat{a}})\leqslant\hat{\eta}(\hat{n}_{\hat{a}},\alpha)\Big]\Big[\,\hat{\eta}(\hat{n}_{\hat{a}},\alpha)\leqslant\eta(\hat{n}_{\hat{a}},\alpha)\Big]\ \leqslant\ \mathbf{P}[\,H_{n_{\hat{a}}}(\hat{n}_{\hat{a}})\leqslant\eta(\hat{n}_{\hat{a}},\alpha)]\ \overset{\text{Th.3.3.3}}{\leqslant}\ \frac{\alpha}{2}.$$

Для выборок второго типа

$$\mathbf{P}\Big[\,H_{n_{\hat{a}}}(\hat{n}_{\hat{a}})\leqslant\hat{\eta}(\hat{n}_{\hat{a}},\alpha)\Big]\Big[\,\hat{\eta}(\hat{n}_{\hat{a}},\alpha)>\eta(\hat{n}_{\hat{a}},\alpha)\Big]\ \overset{1}{\leqslant}\ \mathbf{P}[\,\hat{\eta}(\hat{n}_{\hat{a}},\alpha)>\eta(\hat{n}_{\hat{a}},\alpha)]\ \overset{2}{\leqslant}$$
$$\leqslant\ \sum_{s=0}^{\ell}\mathbf{P}[\,\hat{\eta}(s,\alpha)>\eta(s,\alpha)]\ \overset{3}{=}\ \sum_{s=0}^{\ell}\mathbf{P}\Big[\,2\hat{P}(s,\eta(s,\alpha),\alpha/4\ell)<P(s,\eta(s,\alpha))\Big]\ \overset{4}{\leqslant}\ \frac{\alpha}{2}.$$

Неравенство 2 есть неравенство Буля, равенство 3 следует из того, что P и $\hat{P}$ не убывают по η, неравенство 4 следует из Следствия 4 последней леммы.

Суммируя левые и правые части неравенств для выборок обоих типов, имеем:

$$\mathbf{P}\left[H_{n_{\hat{a}}}(\hat{n}_{\hat{a}}) \leqslant \hat{\eta}(\hat{n}_{\hat{a}}, \alpha)\right] \leqslant \alpha.$$

Используя (1.10) и обращая знак неравенства, получаем утверждение теоремы. ■

К сожалению, профиль $\hat{\Delta}_s$ хотя и основан на обучающей выборке X, является ненаблюдаемой величиной, поскольку множество алгоритмов A, для которого он строится – ненаблюдаемо. При обучении мы имеем в распоряжении лишь *проекцию* A на X:

$$\hat{A} = I(\mathcal{F}, X) = \left\{\left(I(f, x_1), \ldots, I(f, x_\ell)\right) \colon f \in \mathcal{F}\right\},$$

и, соответственно, можем вычислить только профиль наблюдаемых частот этой проекции $\hat{\Delta}_s(\hat{A}) = \operatorname{card}\left\{a \in \hat{A} \colon n(a, X) = s\right\}$. Очевидно, что $\hat{\Delta}_s(\hat{A}) \leqslant \hat{\Delta}_s$.

Для конечного семейства $\mathcal{F}$ (например, для бинарных решающих деревьев), мы можем использовать в последней теореме верхнюю оценку для $\hat{\Delta}_s$ — профиль наблюдаемых частот семейства: $\hat{\Delta}_s(\mathcal{F}) = \operatorname{card}\{f \in \mathcal{F} \colon n(f, X) = s\}$; очевидно, что $\hat{\Delta}_s \leqslant \hat{\Delta}_s(\mathcal{F})$.

В случае бесконечного семейства $\mathcal{F}$ для получения shell-оценки, вычисляемой по X, необходима верхняя оценка профиля $\hat{\Delta}_s$ по профилю $\hat{\Delta}_s(\hat{A})$.

Отметим, что, хотя оценка, полученная в [47], называется наблюдаемой («observable shell bound»), она также может быть вычислена по наблюдаемой выборке только для конечных семейств.

3.5 Выводы

В настоящей главе выводятся комбинаторные аналоги shell-оценок обобщающей способности [49, 47] для случая конечной генеральной совокупности. Комбинаторные оценки не являются асимптотическими и выводятся более общим и простым образом. Показывается, что shell-оценки являются частным случаем (или вариантом) оценок Вапника-Червоненкиса и «бритвы Оккама».

Как показывают эксперименты главы 6, точность shell-оценок несущественно отличается от точности VC-оценки. В эксперименте основная масса алгоритмов в семействе действительно концентрируется в области наихудшей частоты ошибок $\frac{1}{2}$ и при этом метод обучения в основном выбирает алгоритмы из области малых частот

ошибок. То есть, сложность эффективно используемой части семейства существенно меньше сложности всего семейства $\mathcal{F}$. Именно эти факты и приводятся обычно в качестве исходной мотивации shell-оценок. Однако, фактически, в shell-оценках они учитываются не в полной мере. Shell-оценки, как и VC-оценка, основаны на функционале равномерного отклонения частот *по всему* семейству $\mathcal{F}$, а не по его части с малыми частотами ошибок. Основной причиной завышенности по-прежнему остаётся неравенство Буля, в котором суммирование вероятностей производится, опять таки, *по всему* семейству.

Преимущество shell-оценок в том, что они позволяют балансировать точность оценки для разных частот ошибок, делая оценку точнее для одних частот за счет ухудшения оценки для других, аналогично тому, что делается в оценки «бритвы Оккама» для отдельных алгоритмов. Эта идея представляется плодотворной, но выигрыш в точности, который она может дать, полностью нивелируется завышенностью оценки равномерного уклонения и неравенства Буля.

Отметим также, что shell-оценки Теорем 3.3.1 и 3.3.3 получаются из разложения (1.16) оценкой $P(m,s) \leqslant \Delta_m h_m(s)$. Для их уточнения можно было бы рассмотреть оценку равномерного уклонения по одному слою семейства

$$P(m,s) \leqslant \mathbf{P}\Big[\exists a \in A_m \colon \hat{n}_a = s\Big],$$

для правой части которой представляется возможным, основываясь только на свойствах семейства $\mathcal{F}$ и учитывая Хэммингово сходство алгоритмов в слое, получить более точную оценку, чем неравенство Буля.

Глава 4

Оценки на основе характеристик сходства алгоритмов в семействе

4.1 Мотивация и постановка задачи

В VC-оценке (Теорема 1.5.1) функционал равномерного уклонения частот оценивается сверху при помощи неравенства Буля:

$$\mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] \leqslant \sum_{a\in A} \mathbf{P}[\,\mathrm{U}_a]\,. \tag{4.1}$$

Напомним, что здесь U_a есть краткое обозначение условия переобучения алгоритма a. Неравенство Буля достаточно часто используется при получении оценок обобщающей способности в PAC-подходе , хотя известно, что оно сильно завышено при большом числе входящих в него событий.

Неравенство, очевидно, становится точным только если все события в нем взаимоисключающие. При оценивании функционала равномерного уклонения, события U_a наоборот существенно совместны и число событий в неравенстве велико. Как следствие, неравенство Буля представляет собой один из самых существенных факторов завышенности VC-оценки. Рассмотрим пример.

Пример 4.1.1. Пусть размер полной выборки $|\mathbb{X}| = 100$, размер обучающих выборок $\ell = 50$, критерий переобучения $\nu - \hat{\nu} \geqslant \varepsilon$, $\varepsilon = 10\%$.

Тогда, для алгоритма $a \in A$ вероятность того, что он окажется переобучен есть $H_m(s_m(\varepsilon))$ и лежит в диапазоне $0.005 \div 0.05$ (в зависимости от полного числа ошибок m алгоритма).

Пусть $\mathcal{F}$ есть семейство линейных классификаторов в $\mathbb{R}^p$, $p = 10$ и пусть $\mathbb{X}$ находится в общем положении. Тогда полное число алгоритмов в A есть $|A| = 2\sum_{k=0}^{p}\binom{L-1}{k} \approx 3\cdot 10^{13}$. Неравенство Буля (4.1) дает нам $Q_A \leqslant \sum_{a\in A} H_{n_a}(s_{n_a}(\varepsilon)) \approx$

$10^{11} \div 10^{12}$ — оценку для вероятности того, что в A найдется переобученный алгоритм. Очевидно, что эта оценка завышена по крайней мере в 10^{11} раз.

Можно улучшить оценку неравенства Буля учитывая пересечения входящих в него событий. Пример такого учета дает разложение по принципу включения-исключения:

$$Q_A = \mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] = \sum_{S\in 2^A} (-1)^{|S|-1} \mathbf{P}\Big[\bigwedge_{a\in S} \mathrm{U}_a\Big] = \\ = \sum_{a\in A} \mathbf{P}[\,\mathrm{U}_a] \;-\; \sum_{\{a,a'\}\in[A]^2} \mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}] \;+\; \sum_{\{a,a',a''\}\in[A]^3} \mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}\mathrm{U}_{a''}] - \ldots \quad (4.2)$$

Число слагаемых и их сложность (количество U в конъюнкции) в последнем разложении достаточно велико, поэтому на практике пользуются лишь его частью. Отбрасывая в последнем выражении все суммы кроме первой, имеем неравенство Буля. Оставляя таким же образом только первые k сумм для четного/нечетного k, получаем, соответственно, нижнюю/верхнюю оценки для Q_A.

Определение 4.1.2. *Для произвольного конечного множества событий* $\{\mathrm{U}_a\colon a\in A\}$, *индексированного множеством* A, *оценки вероятности дизъюнкции* $\mathbf{P}\Big[\vee_a \mathrm{U}_a\Big]$ *вида*

$$\mathbf{P}\Big[\bigvee_a \mathrm{U}_a\Big] \quad\leqslant\quad \sum_{S\in\mathbb{S}} w_S\, \mathbf{P}\Big[\bigwedge_{a\in S} \mathrm{U}_{a,S}\Big], \qquad \mathbb{S}\subset 2^A, \quad w_S\in\mathbb{R}, \quad \mathrm{U}_{a,S}\in\big\{\mathrm{U}_a, \bar{\mathrm{U}}_a\big\}$$

называются оценками типа Бонферрони.

Обычно стараются минимизировать повторный вклад пересечений событий U_a в различные слагаемые оценки, при этом минимизируя число и сложность слагаемых. Неравенство Буля есть простейшая оценка типа Бонферрони.

Другие примеры точных разложений функционала равномерного уклонения это «цепное» разложение:

$$\mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] \;=\; \sum_{i=1}^{|A|} \mathbf{P}\big[\,\mathrm{U}_{a_i}\bar{\mathrm{U}}_{a_{i-1}}\ldots\bar{\mathrm{U}}_{a_1}\big]$$

и разложение «по подмножествам»:

$$\mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] \;=\; \sum_{S\subset A} \mathbf{P}\Big[\bigwedge_{a\in S} \bar{\mathrm{U}}_a \bigwedge_{a\in A\setminus S} \mathrm{U}_a\Big].$$

При использовании всех этих разложений возникает проблема вычисления вероятностей конъюнкций событий U_a (и их отрицаний) в различных комбинациях.

Очевидно, что пересечение событий $\{\mathrm{U}_a\colon a\in A\}$ определяется сходством алгоритмов в A. Пусть $\forall a, a\in A,\ \rho(a,a') = \operatorname{card}\{x\in\mathbb{X}\colon \mathrm{I}(a,x)\neq\mathrm{I}(a',x)\}$ есть Хэммингово расстояние между алгоритмами a и a'. Тогда на множестве A может быть обычным образом определен неориентированный граф:

Определение 4.1.3. *Графом* 1*-сходства множества* $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ *будем называть граф*

$$\mathrm{G}_A^1 = (A,\, \mathrm{E}), \qquad \mathrm{E} = \{\{a, a'\} \in A \times A \colon \rho(a, a') = 1\},$$

в котором множество ребер E *соединяет алгоритмы, вектора ошибок которых отличаются на одном объекте. Будем называть такие алгоритмы смежными. Для удобства пометим каждое ребро объектом* $x \in \mathbb{X}$*, на котором отличаются алгоритмы соединенные ребром.*

Пример приведен на Рис. 4.1. Отметим, что граф G_A^1 зависит как от семейства $\mathcal{F}$, так и от полной выборки $\mathbb{X}$.

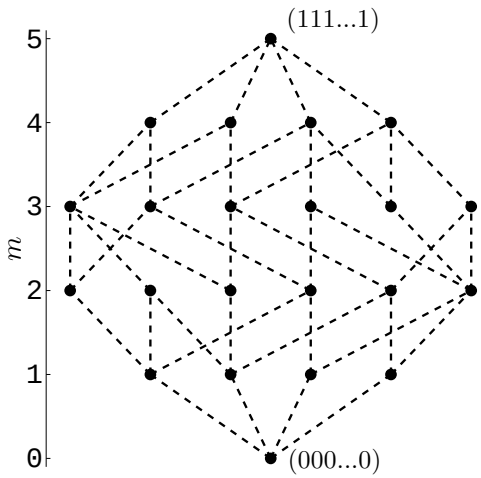

Рис. 4.1. Граф 1-сходства множества из $|A| = 22$ алгоритмов, индуцированного на генеральной совокупности из $|\mathbb{X}| = 5$ объектов в $\mathbb{R}^2$ семейством $\mathcal{F}$ линейных классификаторов.

Горизонтальные слои графа на Рис. 4.1 соответствуют слоям $A_m, m = 0, \ldots, L$ множества A.

Используя представление A в виде графа 1-сходства, можно сказать, что неравенство Буля (4.1) учитывает только число вершин в графе. Цель настоящей главы — учёт метрических свойств графа для получения более точных оценок функционала равномерного уклонения.

4.2 Обзор работ по теме

Понятие графа 1-сходства множества алгоритмов совпадает с понятием графа 1-включения [34]. Однако модель обучения в [34] (the prediction model of learning)

отличается от модели рассмотренной в настоящей статье. Комбинаторные свойства графа 1-включения использовались в [34] для получения оптимальной стратегии прогнозирования значения неизвестной функции из $\mathcal{F}$ на объекте $x \in \mathbb{X}$, если известны значения этой функции на всех объектах $\mathbb{X} \setminus x$. Критерий качества стратегии, для которого в [34] выводится верхняя оценка, это вероятность ошибки на x, в отличие от рассматриваемого нами функционала равномерного уклонения.

Для учета метрических свойств вещественно-значных семейств $\mathcal{F}$ в оценках обобщающей способности, в PAC-подходе широко используется понятие γ-покрытия [43]. При этом неравенство Буля применяется к (конечному) γ-покрытию $\mathcal{F}$ ценой увеличения порога ε на величину порядка γ. Прекрасное введение в методы теории обучения связанные с γ-покрытием дано в [13]; достаточно эффективное применение этот подход нашел в ядерных методах классификации (kernel methods) [21], [35]. Глубокое изложение этих и других методов в теории обучения можно найти в [22]. Отметим здесь также цепной метод Дадли (Dudley's chaining method) [57], позволяющий подбирать γ-покрытие оптимального размера (но все также использующий неравенство Буля).

Проблема получения оценок вероятности конъюнкции событий также возникает в статистике при множественной проверке гипотез (multiple testing) и при выборке по значимости (importance sampling), в теории экстремальных значений, в теории надежности и некоторых проблемах комбинаторики. Несколько методов получения и ряд применений подобных оценок описан в [29]. В [54] для их получения используется подход использующий топологические свойства множества событий входящих в оценку, этот подход развивается в [25] и там же дается его приложение в теории надежности. Особенность этой проблемы в теории обучения в огромном числе событий (по числу алгоритмов в A), входящих в оценку, и отсутствии компактного детерминистского представления структуры их корреляций. В связи с этим, методы получения оценок из работ перечисленных выше, оказываются плохо применимы в теории обучения.

4.3 Вычисление слагаемых в оценках типа Бонферрони

Для подсчета количества выборок в $[\mathbb{X}]^\ell$ удовлетворяющих заданным условиям, достаточно удобен метод производящих функций из перечислительной комбинаторики. Превосходное изложение современных методов аналитической комбинаторики дано

в [28], более краткое изложение с акцентом на производящие функции можно найти в [66].

Пусть $z_i,\ i = 1, \ldots, L$ — формальные переменные, соответствующие объектам полной выборки $x_i \in \mathbb{X}$. Тогда $\mathcal{R} = (1 + z_1)(1 + z_2) \ldots (1 + z_L)$ представляет производящую функцию множества $2^{\mathbb{X}}$ всех подмножеств полной выборки: в разложении $\mathcal{R}$ присутствуют все возможные комбинации переменных z_i с коэффициентом 1.

Пусть все объекты эквивалентны: $z_i \equiv z$, тогда имеем $\mathcal{R} = (1 + z)^L$. Будем обозначать $\left[z^\ell \right] \mathcal{R}$ сумму коэффициентов при z^ℓ в $\mathcal{R}$ после разложения $\mathcal{R}$ на слагаемые. Тогда $\left[z^\ell \right] \mathcal{R} = \left[z^\ell \right] (1 + z)^L = \left[z^\ell \right] \sum_{j=0}^{L} C_L^j z^j = C_L^\ell$ — число подмножеств размера ℓ из L объектов полной выборки.

По аналогии определим операторы $\left[z^{\leqslant k} \right], \left[z^{>k} \right]$, как сумму всех коэффициентов при степенях z, не больших / больших чем k.

Пусть m объектов $\mathbb{X}_m \subset \mathbb{X}$ обладают некоторым свойством, обозначим его a. Сопоставим этим объектам произведение формальных переменных $z_i \equiv za,\ \forall z_i \in \mathbb{X}_m$. Тогда $\mathcal{R} = (1 + z)^{L-m}(1 + za)^m$. Определим $\left[z^\ell a^{\leqslant s} \right] \mathcal{R}$ как сумму коэффициентов при степенях $z^\ell a^k,\ k \leqslant s$ в разложении $\mathcal{R}$. То есть, число подмножеств $\mathbb{X}$, размера ℓ, содержащих не более s объектов, обладающих свойством a.

Пример 4.3.1. Пусть свойство a есть «алгоритм a ошибается на данном объекте», тогда

$$\mathbf{P}[\,\hat{n}_a \leqslant s_{n_a}] \;=\; \tbinom{L}{\ell}^{-1} \left[z^\ell a^{\leqslant s_{n_a}} \right] (1 + z)^{L-m}(1 + za)^m.$$

Имеем

$$\left[z^\ell a^{\leqslant s} \right] (1 + z)^{L-m}(1 + za)^m = \left[z^\ell a^{\leqslant s} \right] \sum_{j=0}^{L-m} C_{L-m}^j z^j \sum_{k=0}^{m} C_m^k z^k a^k =$$

$$= \sum_{j=0}^{L-m} \sum_{k=0}^{m} C_{L-m}^j C_m^k \,[\,j + k = \ell]\,[\,k \leqslant s] = \sum_{i=0}^{s} C_{L-m}^{\ell-i} C_m^i$$

Нормируя на общее число подмножеств размера ℓ имеем гипергеометрическую функцию распределения $H_{L,m}^\ell(s)$.

Пример 4.3.2. Пусть в $\mathbb{X}$ есть объекты двух типов: a и b. К примеру, объекты, на которых ошибаются алгоритмы a и b, соответственно. Пусть m есть число объектов, на которых ошибаются одновременно оба алгоритма. Тогда имеем:

$$\mathbf{P}\,[\,\hat{n}_a \leqslant s_{n_a}]\,[\,\hat{n}_b \leqslant s_{n_b}] \;=$$
$$=\; \tbinom{L}{\ell}^{-1} \left[z^\ell a^{\leqslant s_{n_a}} b^{\leqslant s_{n_b}} \right] (1 + zab)^m (1 + za)^{n_a - m} (1 + zb)^{n_b - m} (1 + z)^{L - n_a - n_b + m}.$$

Раскладывая аналогично предыдущему случаю, несложно показать, что последняя вероятность представляет левый «хвост» заданной формы у 3-мерного гипергеометрического распределения.

Другой удобный подход к определению вероятностей вида $\mathbf{P}[\,\mathrm{U}_{a_1}\ldots\mathrm{U}_{a_k}]$ состоит в следующем.

Определение 4.3.3. *Для произвольного фиксированного упорядочения объектов полной выборки* $\mathbb{X}$, $\hat{\sigma} = (\hat{\sigma}_1,\ldots,\hat{\sigma}_L)^{\mathrm{T}}$ *есть бинарный случайный вектор, в котором* $\forall i, \hat{\sigma}_i \in \{0,1\}$ *есть бинарная случайная переменная, равная 1, если* $x_i \in X$, *и равная* 0 *иначе.*

Для любого вектора $a \in A$ очевидно $\hat{n}_a = a\,\hat{\sigma}$ (здесь скалярное произведение векторов), то есть $[\,\mathrm{U}_a] \;=\; [\,a\,\hat{\sigma} \leqslant s_{n_a}]$. Пусть пусть $a_1,\ldots,a_k$ — какие-то вектора ошибок из A с $m_1,\ldots,m_k$ ошибками на полной выборке, соответственно. Имеем

$$\mathbf{P}[\,\mathrm{U}_{a_1}\ldots\mathrm{U}_{a_k}] \;=\; \mathbf{P}\,[\,a_1\,\hat{\sigma} \leqslant s_{m_1}]\ldots[\,a_k\,\hat{\sigma} \leqslant s_{m_k}]\,.$$

Определение 4.3.4. *Для векторов* $a_1,\ldots,a_k$ *из* A, $\mathbf{M}(a_1,\ldots,a_k)$ *есть бинарная* $k\times$ $\times\, L$ *матрица в которой вектор* a_i *представляет* i*-ую строку.* $\mathrm{M}_{ij} = 1$, *если и только если алгоритм* a_i *делает ошибку на объекте* x_j. *Каждая обучающая выборка* X *соответствует выбору некоторых* ℓ *столбцов из* $\mathbf{M}$.

Обозначим $\mathbf{s} = (s_{m_1},\ldots,s_{m_k})^{\mathrm{T}}$ вектор пороговых значений числа ошибок на обучающей выборке для алгоритмов $a_1,\ldots,a_k$. Тогда набор условий $\mathrm{U}_{a_1}\ldots\mathrm{U}_{a_k}$ записывается как $\mathbf{M}\hat{\sigma} \leqslant \mathbf{s}$ и, соответственно, $\mathbf{P}[\,\mathrm{U}_{a_1}\ldots\mathrm{U}_{a_k}]$ есть доля выборок, для которых выполняется $\mathbf{M}\hat{\sigma} \leqslant \mathbf{s}$. Будем кратко записывать

$$\mathbf{P}\Big[\,\mathbf{M}\hat{\sigma} \leqslant \mathbf{s}\Big] \;\stackrel{\text{def}}{=}\; \mathbf{P}\,[\,a_1\,\hat{\sigma} \leqslant s_{m_1}]\ldots[\,a_k\,\hat{\sigma} \leqslant s_{m_k}]\,.$$

Если вместо части условий U_a в искомую вероятность входят их отрицания $\bar{\mathrm{U}}_a$, то знаки соответствующих неравенств в системе меняются на обратные.

Пример 4.3.5. Рассмотрим два алгоритма $a, a' \in A$ таких, что $\rho(a,a') = 1,\ n_a = m$, $n_{a'} = m+1$. Тогда система $\mathbf{M}\hat{\sigma} \leqslant \mathbf{s}$ есть:

$$\begin{cases}\mathrm{U}_a\\ \mathrm{U}_{a'}\end{cases} \Leftrightarrow \begin{cases}(11\ldots1\mathbf{0}00\ldots0)\cdot\hat{\sigma} \;\leqslant\; s_m\\ (11\ldots1\mathbf{1}00\ldots0)\cdot\hat{\sigma} \;\leqslant\; s_{m+1}\end{cases} \Leftrightarrow \begin{cases}\hat{n}_a & \leqslant\; s_m\\ \hat{n}_a + \hat{\sigma}_{m+1} & \leqslant\; s_{m+1}\end{cases}.$$

При $s_{m+1} = s_m$ несложно видеть, что если первое условие не выполняется, то второе также не выполняется, то есть система эквивалентна одному условию $\hat{n}_a \leqslant s_m$ и $\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}] = H_m(s_m)$.

Между записью вероятности $\mathbf{P}[\,\mathrm{U}_{a_1}\ldots\mathrm{U}_{a_k}]$ через производящий многочлен и через систему линейных неравенств легко устанавливается соответствие.

Утверждение 4.3.6. *Доля выборок, удовлетворяющих системе* $\mathbf{M}\hat{\sigma}\leqslant\mathbf{s}$ *с* k *условиями, есть*

$$\mathbf{P}\Big[\,\mathbf{M}\hat{\sigma}\leqslant\mathbf{s}\Big]\;=\;\big[\,z^{\ell}a_1^{\leqslant \mathrm{s}_1}\ldots a_k^{\leqslant \mathrm{s}_k}\big]\prod_{i=1}^{L}\big(1+z\,a_1{}^{\mathrm{M}_{1i}}\ldots a_k{}^{\mathrm{M}_{ki}}\big)\,.$$

Последнее также справедливо при замене в любом условии системы знака $\leqslant$ *на любой из* $\{\leqslant,\geqslant,<,>,=\}$ *с заменой соответствующего знака в правой части равенства.*

4.4 Оценка с учетом связности семейства

Предположим, что граф 1-сходства множества A связен. Это предположение выполняется для многих семейств $\mathcal{F}$, непрерывных по параметрам, в которых любой алгоритм может быть преобразован в любой другой алгоритм из $A=\mathrm{I}(\mathcal{F},\mathbb{X})$ путём изменения вектора параметров вдоль некоторой непрерывной траектории в пространстве параметров так, что соответствующий вектор ошибок меняется каждый раз на одном объекте. В частности, в [61] показано, что если $\mathbb{X}$ взята из некоторого распределения и $\mathcal{F}$ непрерывно по параметрам, то при достаточно слабых условиях на гладкость распределения и $\mathcal{F}$, выполняющихся для многих практических семейств, граф 1-сходства связен с вероятностью 1.

Рассмотрим два смежных алгоритма $a,a'\in A$. Неравенство Буля для них есть $\mathbf{P}[\,\mathrm{U}_a\vee\mathrm{U}_{a'}]\leqslant\mathbf{P}[\,\mathrm{U}_a]+\mathbf{P}[\,\mathrm{U}_{a'}]$, оно завышено в точности на $\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}]$. Поскольку a и a' отличаются на 1 бит, можно предположить, что вероятность $\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}]$ близка по величине к $\mathbf{P}[\,\mathrm{U}_a]$ и $\mathbf{P}[\,\mathrm{U}_{a'}]$. Действительно, имеем следующую лемму.

Лемма 4.4.1. *Для любого семейства функций* $\mathcal{F}$*, полной выборки* $\mathbb{X}$*,* $|\mathbb{X}|=L$*, индикатора ошибки* I*,* $\forall\varepsilon\in(0,1)$ *и* $A=\mathrm{I}(\mathcal{F},\mathbb{X})$*, пусть* $a,a'\in A$ *такие, что* $\rho(a,a')=1$*,* $n_a=m,\;\;n_{a'}=m+1$*. Тогда*

$$\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}]=\begin{cases}\mathbf{P}[\,\mathrm{U}_a]\,, & \text{если } s_m<s_{m+1};\\ \mathbf{P}[\,\mathrm{U}_{a'}]\,, & \text{если } s_m=s_{m+1};\end{cases}$$

где s_m *произвольная граница переобучения.*

Доказательство. Пусть $\hat{\sigma}_{m+1}$ соответствует объекту, на котором отличаются a и a'.

Тогда очевидно $\hat{n}_{a'}=\hat{n}_a+\hat{\sigma}_{m+1}$ и $\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}]=\mathbf{P}\,[\,\hat{n}_a\leqslant s_m]\,[\,\hat{n}_a+\hat{\sigma}_{m+1}\leqslant s_{m+1}]$. Легко видеть, что при $s_m<s_{m+1}$ второе условие является следствием первого, то

есть $\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}] = \mathbf{P}[\,\hat{n}_a \leqslant s_m] = \mathbf{P}[\,\mathrm{U}_a]$. Аналогично, при $s_m = s_{m+1}$ первое условие является следствием второго и $\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}] = \mathbf{P}[\,\mathrm{U}_{a'}]$.

Лемма доказана. ■

Если граф множества A связный, то у каждого алгоритма в A найдется по крайней мере один смежный с ним алгоритм и можно уточнить неравенство Буля для A, вычтя попарные вероятности переобучения для смежных алгоритмов аналогично тому, как это сделано выше. В частности, имеет место следующая оценка типа Бонферрони.

Лемма 4.4.2 (Hunter [37]). *Пусть* $\{\mathrm{U}_a\colon a \in A\}$ *есть произвольное конечное множество событий, индексированных множеством* A. *Пусть* $(A,\ \mathrm{T} \subset A \times A)$ *есть произвольное неориентированное дерево на* A. *Тогда выполняется верхняя оценка:*

$$\mathbf{P}\Big[\bigvee_{a\in A}\mathrm{U}_a\Big] \leqslant \sum_{a\in A}\mathbf{P}[\,\mathrm{U}_a] \;-\; \sum_{\{a,a'\}\in T}\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}]\,.$$

В связном графе 1-сходства всегда можно выделить подграф-дерево. Тогда имеем следующую оценку обобщающей способности. Здесь используется обычный квантильный функционал вероятности переобучения Q_η.

Теорема 4.4.3. *Для любого семейства* $\mathcal{F}$, *любой полной выборки* $\mathbb{X}$, *метода обучения* $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$ *и индикатора ошибки* I, *если граф* G^1_A *множества* $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ *связный, то выполняется оценка вероятности переобучения:*

$$\mathbf{P}[\,n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta)] \;\leqslant\; P_{tree}(\eta) \;\stackrel{\text{def}}{=}\; \eta + |A| \max_m h_m(s_m(\eta))\,.$$

Также верна доверительная оценка

$$\mathbf{P}[\,n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{tree}(\alpha))] \;\leqslant\; \alpha,$$

где $\eta_{tree}(\alpha) \;=\; \max\{\eta\colon P_{tree}(\eta) \leqslant \alpha\}$.

Доказательство. Выделим в графе G^1_A произвольное дерево (A, T). Выберем произвольный алгоритм a_0 в качестве его вершины и ориентируем все рёбра по направлению к вершине; тогда имеем ориентированное дерево (A, T'). Поставим в соответствие каждому алгоритму в A (кроме корня) ребро, исходящее из него. Тогда из Леммы 4.4.2 имеем:

$$\mathbf{P}\Big[\bigvee_{a\in A}\mathrm{U}_a\Big] \;\leqslant\; \sum_{a\in A}\mathbf{P}[\,\mathrm{U}_a] - \sum_{\{a,a'\}\in \mathrm{T}}\mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}] \;=\; \mathbf{P}[\,\mathrm{U}_{a_0}] + \sum_{(a,a')\in \mathrm{T}'}(\mathbf{P}[\,\mathrm{U}_a] - \mathbf{P}[\,\mathrm{U}_a\mathrm{U}_{a'}])\,.$$

В соответствии с Леммой 4.4.1 некоторые из слагаемых в последней сумме равны 0, остальные равны $\mathbf{P}[\,\mathrm{U}_a] - \mathbf{P}[\,\mathrm{U}_{a'}]$. Обозначим для краткости $n_a = n,\ n_{a'} = n'$. Тогда:

$$\mathbf{P}[\,\mathrm{U}_a] - \mathbf{P}[\,\mathrm{U}_{a'}] = \begin{cases} 0, & \text{при } n' = n+1, \quad s_{n'} = s_n + 1; \\ 0, & \text{при } n' = n-1, \quad s_{n'} = s_n; \\ H_n(s_n) - H_{n+1}(s_n)\,, & \text{при } n' = n+1, \quad s_{n'} = s_n; \\ H_n(s_n) - H_{n-1}(s_n - 1)\,, & \text{при } n' = n-1, \quad s_n = s_{n'} + 1. \end{cases}$$

Два последних варианта оцениваются сверху как $h_n(s_n)$ исходя из свойств гипергеометрического распределения (Утверждение 1.3.3). То есть имеем $\mathbf{P}[\,\mathrm{U}_a] - \mathbf{P}[\,\mathrm{U}_{a'}] \leqslant h_n(s_n)$. Подставляя это в последнее выражение, беря максимум по n и обращая оценку имеем утверждение теоремы. ■

В сравнении с VC-оценкой Теоремы 1.5.1 имеем в последней теореме замену левого «хвоста» гипергеометрического распределения $H_m(s_m(\eta))\,(\approx \eta)$ на значение гипергеометрической вероятности в точке $h_m(s_m(\eta))$. К сожалению, эта величина оказывается не достаточно мала в сравнении с η, что дает в итоге оценку, лишь немного более точную, чем VC-оценка. Отметим, что связность графа на множестве алгоритмов также использовалась для получения оценок обобщающей способности в [61] и, хотя применялся другой метод, так же без существенного улучшения оценок.

По сути, учет факта связности графа 1-сходства означает учёт наличия одной связи у каждого алгоритма в A. Для получения более точной оценки необходим более глубокий учет связности графа.

4.5 Оценка с учетом распределения полустепеней связности алгоритмов

Определение 4.5.1. *Верхняя/нижняя 1-окрестность алгоритма $a \in A$ в графе 1-сходства есть множество смежных с a алгоритмов в следующем / предыдущем слое множества A:*

$$A_{\pm}(a) = \{a' \in A\colon \rho(a,a') = 1,\ n(a',\mathbb{X}) = n(a,\mathbb{X}) \pm 1\}\,.$$

Определение 4.5.2. *Множество верхних/нижних связующих объектов алгоритма a есть множество объектов, на которых a отличается от алгоритмов своей верхней / нижней 1-окрестности:*

$$\mathbb{X}_{\pm}(a) = \{x \in \mathbb{X}\colon \exists a' \in A_{\pm}(a),\ \mathrm{I}(a,x) \neq \mathrm{I}(a',x)\}\,.$$

Определение 4.5.3. *Верхняя / нижняя полустепень связности алгоритма a есть есть число алгоритмов в его верхней / нижней единичной окрестности:*

$$\rho_{\pm}(a) = |A_{\pm}(a)| = |\mathbb{X}_{\pm}(a)|\,.$$

Рассмотрим цепное разложение функционала равномерного уклонения (для некоторого фиксированного упорядочения алгоритмов в A):

$$\mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] = \mathbf{P}[\,\mathrm{U}_{a_1}] + \mathbf{P}\big[\,\mathrm{U}_{a_2}\bar{\mathrm{U}}_{a_1}\big] + \mathbf{P}\big[\,\mathrm{U}_{a_3}\bar{\mathrm{U}}_{a_2}\bar{\mathrm{U}}_{a_1}\big] + \ldots \tag{4.3}$$

Напомним, что отрицание $\bar{\mathrm{U}}_a$ здесь есть краткое обозначение для $n(a,X) > s_{n(a,\mathbb{X})}$, и имеет смысл "алгоритм a не переобучен на X". Отметим, что отбрасывая в последнем разложении все $\bar{\mathrm{U}}$-условия, мы получаем неравенство Буля.

Рассмотрим j-ый член разложения: $\mathbf{P}\big[\,\mathrm{U}_{a_j}\bar{\mathrm{U}}_{a_{j-1}}\bar{\mathrm{U}}_{a_{j-2}}\ldots\big]$. Если среди $\bar{\mathrm{U}}$-условий есть условия, соответствующие алгоритмам из 1-окрестности $A_{\pm}(a_j)$, то они, в некотором смысле, противоречат условию U_{a_j}; как следствие, последняя вероятность должна быть мала по сравнению с $\mathbf{P}\big[\,\mathrm{U}_{a_j}\big]$. Отбрасывая все $\bar{\mathrm{U}}$-условия, кроме соответствующих алгоритмам из $A_{\pm}(a_j)$, мы, очевидно, получаем верхнюю оценку для j-ого члена разложения. Точное значение этой оценки дается следующей леммой.

Лемма 4.5.4. *Для любого семейства функций $\mathcal{F}$, полной выборки $\mathbb{X}, |\mathbb{X}| = L$, индикатора ошибки* I*, функции границы переобучения s_m, пусть $A = \mathrm{I}(\mathcal{F},\mathbb{X})$, алгоритм $a \in A$ имеет полное число ошибок $n(a,\mathbb{X}) = m$ и алгоритмы $\{b_1,\ldots,b_q\} \subseteq A_+(a)$, $\{c_1,\ldots,c_p\} \subseteq A_-(a)$, $p+q>0$ находятся в верхней/нижней 1-окрестности a, тогда:*

$$\mathbf{P}\big[\,\mathrm{U}_a\bar{\mathrm{U}}_{b_1}\ldots\bar{\mathrm{U}}_{b_q}\bar{\mathrm{U}}_{c_1}\ldots\bar{\mathrm{U}}_{c_p}\big] \;=\; \begin{cases} 0, & \text{если } s_{m-1} = s_m,\ p \neq 0; \\ 0, & \text{если } s_{m+1} = s_m + 1,\ q \neq 0; \\ \dfrac{\binom{m-p}{s_m}\binom{L-m-q}{\ell-s_m-q}}{\binom{L}{\ell}}, & \text{иначе.} \end{cases}$$

Доказательство. Набор условий $\mathrm{U}_a\bar{\mathrm{U}}_{b_1}\ldots\bar{\mathrm{U}}_{b_q}\bar{\mathrm{U}}_{c_1}\ldots\bar{\mathrm{U}}_{c_p}$ записывается как:

$$\begin{cases} (111\ldots1\,000\ldots0)\cdot\hat{\sigma} & \leqslant s_m \\ (111\ldots1\,\mathbf{1}00\ldots0)\cdot\hat{\sigma} & > s_{m+1} \\ (111\ldots1\,0\mathbf{1}0\ldots0)\cdot\hat{\sigma} & > s_{m+1} \\ & \ldots \\ (\mathbf{0}11\ldots1\,000\ldots0)\cdot\hat{\sigma} & > s_{m-1} \\ (1\mathbf{0}1\ldots1\,000\ldots0)\cdot\hat{\sigma} & > s_{m-1} \\ & \ldots \end{cases} \Leftrightarrow \begin{cases} \hat{n}_a & \leqslant s_m \\ \hat{n}_a + \hat{\sigma}_{m+1} & > s_{m+1} \\ \hat{n}_a + \hat{\sigma}_{m+2} & > s_{m+1} \\ & \ldots \\ \hat{n}_a - \hat{\sigma}_1 & > s_{m-1} \\ \hat{n}_a - \hat{\sigma}_2 & > s_{m-1} \\ & \ldots \end{cases} \Leftrightarrow \begin{cases} \hat{n}_a & = s_m \\ \hat{\sigma}_{m+1} & = 1 \\ \hat{\sigma}_{m+2} & = 1 \\ & \ldots \\ \hat{\sigma}_1 & = 0 \\ \hat{\sigma}_2 & = 0 \\ & \ldots \end{cases}$$

Здесь первая строка соответствует алгоритму a, следующие q строк — алгоритмам $\{b_1, \ldots, b_q\}$ и последние p строк — алгоритмам $\{c_1, \ldots, c_p\}$. Жирным отмечены позиции, в которых алгоритмы отличаются от a. Легко видеть, что исходная система эквивалентна следующей за ней.

Далее, из второй системы легко видеть, что если $s_{m+1} > s_m$ и $q \neq 0$, то первое условие несовместно с q условиями b-алгоритмов; а также, если $s_{m-1} = s_m$ и $p \neq 0$ то первое условие несовместно с p условиями c-алгоритмов. В обоих этих случаях искомая вероятность равна нулю. Пусть $s_{m-1} = s_m - 1$ при $p > 0$ и $s_{m+1} = s_m$ при $q > 0$. Тогда очевидна последняя эквивалентность.

Число выборок из L по ℓ таких, что из первых m объектов выбрано ровно s_m, но при этом не выбран ни один из первых p объектов, и выбраны все объекты с $(m+1)$-го по $(m+q)$-ый, очевидно, есть $\binom{m-p}{s_m}\binom{L-m-q}{\ell-s_m-q}$. Для доказательства можно воспользоваться Утверждением 4.3.6 для перехода от последней системы к производящему полиному для таких выборок. Нормируя на $\binom{L}{\ell}$ имеем утверждение леммы. ■

Отметим, что искомая вероятность:

$$\mathbf{P}\big[\,\mathrm{U}_{a_j}\bar{\mathrm{U}}_{b_1}\ldots\bar{\mathrm{U}}_{b_q}\bar{\mathrm{U}}_{c_1}\ldots\bar{\mathrm{U}}_{c_p}\big] \;=\; \mathbf{P}\big[\,\mathrm{U}_{a_j}\big] - \mathbf{P}\big[\,\mathrm{U}_{a_j}\big(\mathrm{U}_{b_1}\vee\cdots\vee\mathrm{U}_{b_q}\vee\mathrm{U}_{c_1}\vee\cdots\vee\mathrm{U}_{c_p}\big)\big]\,,$$

по сути есть член $\mathbf{P}\big[\,\mathrm{U}_{a_j}\big]$ неравенства Буля за вычетом вероятности одновременного переобучения a_j и любого из смежных с ним алгоритмов. Мы вычитаем последнюю вероятность, поскольку она учитывается многократно в различных членах неравенства Буля; такое вычитание представляет собой форму принципа включения-исключения.

Для удобства представим последнюю оценку в следующем виде.

Лемма 4.5.5.

$$\frac{\binom{m-p}{s_m}\binom{L-m-q}{\ell-s_m-q}}{\binom{L}{\ell}} \;\leqslant\; \big(\tfrac{\ell-s_m}{L-m}\big)^q\,\big(\tfrac{m-s_m}{m}\big)^p\,h_m(s_m)\,, \qquad \textit{где}\quad \tfrac{\ell-s_m}{L-m}<1,\quad \tfrac{m-s_m}{m}<1.$$

Доказательство. Имеем:

$$\tbinom{m-p}{s_m}\tbinom{L-m-q}{\ell-s_m-q} \;=\; \tfrac{(m-s_m-p+1)\ldots m-s_m}{(m-p+1)\ldots m}\tbinom{m-p}{s_m}\;\tfrac{(\ell-s_m-q+1)\ldots(\ell-s_m)}{(L-m-q+1)\ldots(L-m)}\tbinom{L-m}{\ell-s_m}.$$

Оценивая при $i > 0$

$$\tfrac{\ell-s_m-i+1}{L-m-i+1} \leqslant \tfrac{\ell-s_m}{L-m}$$

и

$$\tfrac{m-s_m-i+1}{m-i+1} \leqslant \tfrac{m-s_m}{m}$$

и пользуясь определением $h_m(s) = \binom{m}{s}\binom{L-m}{\ell-s}\binom{L}{\ell}^{-1}$ имеем утверждение леммы. ■

Последняя лемма показывает, что оценка Леммы 4.5.4 уменьшается экспоненциально с ростом p и q — числа учтённых алгоритмов в 1-окрестности a.

Этот факт имеет простую интерпретацию. Как видно из доказательства Леммы 4.5.4, каждое $\bar{\mathrm{U}}$-условие приводит к «фиксации» состояния $\hat{\sigma}_i$ какого-то объекта $x_i \in \mathbb{X}$, (то есть x_i должен принадлежать или наоборот не принадлежать обучающей выборке, чтобы это условие выполнялось). Такая фиксация уменьшает количество допустимых (удовлетворяющих всем условиям) обучающих выборок в некоторое число раз. К примеру, если рассмотреть все возможные 2^L подмножеств $\mathbb{X}$, то фиксация состояния q объектов приводит к уменьшению числа возможных подмножеств в 2^q раз.

Упорядочим алгоритмы в A по убыванию их полного числа ошибок, то есть «сверху вниз» в графе 1-сходства. Порядок алгоритмов внутри одного слоя несущественнен. Тогда легко видеть, в разложении (4.3) отдельный член оценивается Леммой 4.5.4 с $p = 0$ и $q = \rho_+(a_j)$.

Определение 4.5.6. *Профиль расслоения-связности множества* $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ *есть матрица чисел:*

$$\Delta_{m,q} \;=\; \operatorname{card}\{a \in A\colon n(a,\mathbb{X}) = m,\ \rho_+(a) = q\},$$

где $m = 0, \ldots, L$, $q = 0, \ldots, \max\limits_a \rho_+(a)$. $\Delta_{m,q}$ *есть число алгоритмов в* m*-ом слое, имеющих ровно* q *смежных алгоритмов в* $m{+}1$*-ом слое.*

Если рассматривать равновероятный выбор алгоритма из A, то $\Delta_{m,q}$ есть дискретное совместное распределение величин n_a и $\rho_+(a)$. Его частные распределения это профиль расслоения A:

$$\Delta_m \;=\; \operatorname{card}\{a \in A\colon n(a,\mathbb{X}) = m\} \;=\; \sum_{q=0}^{L} \Delta_{m,q}$$

и *профиль связности* A:

$$\Delta_q^+ \;=\; \operatorname{card}\{a \in A\colon \rho_+(a) = q\} \;=\; \sum_{m=0}^{L} \Delta_{m,q}\,.$$

Имеем следующую оценку обобщающей способности.

Теорема 4.5.7. *Для любого семейства* $\mathcal{F}$*, полной выборки* $\mathbb{X}$*, индикатора ошибки* I*, если профиль расслоения-связности множества* $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ *есть* $\Delta_{m,q}$*, то для любого метода обучения* $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$ *вероятность переобучения оценивается как*

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta)] \;\leqslant\; P_{conn}(\eta) \;\stackrel{\text{def}}{=}\; \eta N_0 \;+\; \tilde{\eta} \sum_{q=1}^{L} \sum_{m=0}^{L} \Delta_{m,q} \alpha_m^q,$$

где N_0 — число алгоритмов в A с пустой верхней 1-окрестностью, $\tilde{\eta} = \max\limits_m h_m(s_m)$, $\alpha_m = \frac{\ell - s_m}{L-m}\left[\frac{\ell-s_m}{L-m} < 1\right]$. Также верна доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{conn}(\alpha))] \;\leqslant\; \alpha, \tag{4.4}$$

где $\eta_{conn}(\alpha) \;=\; \max\{\eta\colon P_{conn}(\eta) \leqslant \alpha\}$.

Доказательство. Упорядочим алгоритмы в A по убыванию полного числа ошибок, тогда из разложения (4.3) имеем

$$\mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] \;=\; \sum_{i=1}^{|A|} \mathbf{P}\big[\,\mathrm{U}_{a_i}\bar{\mathrm{U}}_{a_{i-1}}\ldots\bar{\mathrm{U}}_1\big] \;\leqslant\; \sum_{i=1}^{|A|} \mathbf{P}\big[\mathrm{U}_{a_i} \bigwedge_{a\in A_+(a_i)} \bar{\mathrm{U}}_a\big]. \tag{4.5}$$

Рассмотрим i-ый член последней суммы, обозначим его для краткости P_i. Обозначим также для краткости $n_{a_i} = m$ и $\rho_+(a_i) = q$. Рассмотрим 2 случая.

Пусть $A_+(a_i) \neq \varnothing$. Из Лемм 4.5.4 и 4.5.5 имеем $P_i \;\leqslant\; \left(\frac{\ell-s_m}{L-m}\right)^q$. Кроме того, заметим, что при $\frac{\ell-s_m(\eta)}{L-m} > 1$ мы имеем в Лемме 4.5.4 $\binom{L-m-q}{\ell-s_m(\eta)-q} = 0$ и тогда $P_i = 0$.

Пусть $A_+(a_i) = \varnothing$. Тогда $P_i = \mathbf{P}[\,\mathrm{U}_{a_i}] \leqslant \eta$.

Подставляя полученные оценки в (4.5), используя определение профиля $\Delta_{m,q}$ и обращая оценку имеем утверждение теоремы. ■

Отметим, что теорема записана для квантильного критерия переобучения, но аналогичная оценка выполняется и для критерия $\nu - \hat{\nu} \geqslant \varepsilon$ с заменой в Теореме $\bar{n}\,(\hat{n}_{\hat{a}}, \eta) \;\to\; L(\hat{n}_{\hat{a}}/\ell + \varepsilon)$ и $\eta \;\to\; P_1(\varepsilon)$.
Заметим, что VC-оценка может быть записана как:

$$\mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] \;\leqslant\; \eta\,|A| \;=\; \eta \sum_{m,q} \Delta_{m,q}$$

Таким образом, в сравнении с VC-оценкой, в теореме имеем замену гипергеометрического «хвоста» η на гипергеометрическую вероятность $\tilde{\eta}$ и множитель α_m^q, где q верхняя полустепень связности алгоритма. Величина $\sum_m \Delta_{m,q}\alpha_m^q$ в этом случае есть эффективный аналог коэффициента разбиения $|A|$ в VC-оценке. К примеру, если $L = 100, \ell = 50, \varepsilon = 0.05$, то имеем $\alpha_m \approx \frac{1}{2}$ и, соответственно, множители 2^{-q} к слагаемым VC-оценки. В эксперименте с семейством линейных классификаторов главы 6, число N_0 алгоритмов без верхних связей в A пренебрежимо мало́ в сравнении с общим числом алгоритмов, что дает экспоненциальное улучшение оценки последней теоремы относительно VC-оценки с ростом среднего q. Среднее число связей q, в свою очередь, растет с ростом размерности, вплоть до $A = \{0,1\}^L$. Интуитивно представляется, что число алгоритмов без верхних связей должно быть мало́ также и для других практически используемых семейств $\mathcal{F}$ и индикаторов ошибки I.

Пример 4.5.8. Рассмотрим множество A, $|A| = 2^L$ представляющее полный булев куб, что соответствует семейству $\mathcal{F}$ с $\mathrm{VCdim}(\mathrm{I}(\mathcal{F})) > L$. Пусть критерий переобучения $\nu_a - \hat{\nu}_a \geqslant \varepsilon$, $\varepsilon = 5\%$, то есть $s_m = m\frac{\ell}{L} - \varepsilon\ell$, возьмём для простоты $\ell = L/2$. VC-оценка есть

$$\mathbf{P}[\,\mathrm{U}_{\hat{a}}] \;\leqslant\; P_1(\varepsilon) \cdot 2^L.$$

Профиль расслоения-связности булева куба есть $\Delta_{m,q} = C_L^m\,[\,q = L - m]$. Оценка последней теоремы записывается как

$$\mathbf{P}[\,\mathrm{U}_{\hat{a}}] \;\leqslant\; P_1(\varepsilon) + \tilde{\eta} \sum_{m=0}^{L} C_L^m \left(\tfrac{\ell - s_m}{L-m}\right)^{L-m} \;\leqslant\; P_1(\varepsilon) \cdot 2^{0.6L},$$

то есть учет связности уменьшает оценку вероятности переобучения примерно в $2^{0.4L}$ раз. В общем случае, A представляет подмножество булева куба и уменьшение оценки за счет учета степени связности может быть как более, так и менее существенным, чем для полного булева куба.

Отметим, что точная форма профиля $\Delta_{m,q}$ множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ зависит от полной выборки $\mathbb{X}$ как и от семейства $\mathcal{F}$. Таким образом, последняя оценка в терминах PAC-подхода является "зависящей от распределения" (distribution dependent) и для практического применения ее необходимо связать с наблюдаемой выборкой X и свойствами семейства $\mathcal{F}$. Первый шаг в этом направлении делается в главе 5, где для семейства классификаторов, линейных по параметрам, выводятся оценки среднего и дисперсии профиля расслоения, не зависящие от полной выборки.

Далее, заметим, что оценка настоящего параграфа учитывает для каждого $a \in A$ только алгоритмы смежные с ним и не принимает во внимание алгоритмы находящиеся от a на расстоянии больше единицы. Рассмотрим простейшую структуру содержащую алгоритмы на различных расстояниях от a — монотонную цепь алгоритмов и определим какое улучшение ее учет может дать в оценке члена цепного разложения (4.3).

4.6 Оценка с учетом монотонных цепей алгоритмов

Определение 4.6.1. *Цепь алгоритмов есть упорядоченный набор алгоритмов $a_1, \ldots, a_K$, каждый из которых отличается от предшествующего на 1 произвольном объекте $\rho(a_k, a_{k-1}) = 1$. Будем называть цепь монотонной, если каждый алгоритм в цепи допускает на 1 ошибку больше, чем предшествующий $n_{a_k} = n_{a_{k-1}} + 1$.*

Наличие цепей в A достаточно естественное предположение для $\mathcal{F}$ с непрерывными параметрами; цепь может возникать, если выбирается некоторая «начальная» функция в $\mathcal{F}$ и ее параметры изменяются вдоль некоторого непрерывного пути в пространстве параметров.

Используем снова разложение (4.3) функционала равномерного уклонения и то же упорядочение алгоритмов (от верхнего уровня к нижнему), которое было использовано в предыдущем параграфе. Рассмотрим i-ое слагаемое разложения (4.3), $\mathbf{P}\big[\,\mathrm{U}_i\bar{\mathrm{U}}_{i-1}\ldots\bar{\mathrm{U}}_1\big]$. Предположим, из алгоритма a_i исходит некоторая монотонная цепь алгоритмов, тогда каждому алгоритму цепи соответствует $\bar{\mathrm{U}}$-условие в последней вероятности. Отбрасывая в последней вероятности все $\bar{\mathrm{U}}$-условия кроме соответствующих алгоритмам цепи, мы получаем верхнюю оценку для i-го члена разложения. Точное значение этой оценки дается следующей леммой.

Лемма 4.6.2. *Для любого семейства функций $\mathcal{F}$, полной выборки $\mathbb{X}$, $|\mathbb{X}| = L$, индикатора ошибки* I, *критерия переобучения* U*; пусть $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ и $a \in A$ есть алгоритм с полным числом ошибок $n(a, \mathbb{X}) = m$ и исходящей из него монотонной цепью алгоритмов $(a, b_1, \ldots, b_K)$, $K > 0$ такой, что $n(b_i, \mathbb{X}) = m + i$. Тогда*

$$\mathbf{P}\big[\,\mathrm{U}_a\bar{\mathrm{U}}_{b_1}\ldots\bar{\mathrm{U}}_{b_K}\big] = \frac{\binom{m}{s_m}\binom{L-m}{\ell-s_m}_{\boldsymbol{u}}}{\binom{L}{\ell}},$$

где $\binom{L-m}{\ell-s_m}_{\boldsymbol{u}}$ усечённый биномиальный коэффициент, определяемый рекуррентным отношением

$$\binom{n}{k}_{\boldsymbol{u}} = \left(\binom{n-1}{k}_{\boldsymbol{u}} + \binom{n-1}{k-1}_{\boldsymbol{u}}\right) \cdot [\,k > u_n\,];$$

с граничными условиями $\binom{n}{0}_{\boldsymbol{u}} = 1$ и $\boldsymbol{u}$ есть вектор ограничений длины $(L-m)$,

$$\boldsymbol{u} = \Big((s_{m+1} - s_m),\ \ldots,\ (s_{m+K} - s_m),\ \ldots,\ (s_{m+K} - s_m)\Big),$$

соответствующих условиям $\bar{\mathrm{U}}_{b_1}\ldots\bar{\mathrm{U}}_{b_K}$.

Мы определили здесь для удобства новую комбинаторную величину — усечённый биномиальный коэффициент. Напомним, что обычный коэффициент определяется соотношением $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$.

Доказательство. Искомая вероятность есть

$$\mathbf{P}\big[\,\mathrm{U}_a\bar{\mathrm{U}}_{b_1}\ldots\bar{\mathrm{U}}_{b_K}\big] \;=\; \mathbf{P}\,[\,a\hat{\sigma} \leqslant s_m]\,[\,b_1\hat{\sigma} > s_{m+1}]\ldots[\,b_K\hat{\sigma} > s_{m+K}].$$

Последний набор условий записывается как:

$$\begin{cases} (1\ldots1\,0000\ldots0)\cdot\hat\sigma \ \leqslant s_m \\ (1\ldots1\,\mathbf{1}000\ldots0)\cdot\hat\sigma \ > s_{m+1} \\ (1\ldots1\,\mathbf{11}00\ldots0)\cdot\hat\sigma \ > s_{m+2} \\ (1\ldots1\,\mathbf{111}0\ldots0)\cdot\hat\sigma \ > s_{m+3} \\ \quad\ldots \end{cases} \Leftrightarrow \begin{cases} \hat n_a & \leqslant s_m \\ \hat n_a + \hat\sigma_{m+1} & > s_{m+1} \\ \hat n_a + \hat\sigma_{m+1} + \hat\sigma_{m+2} & > s_{m+2} \\ \hat n_a + \hat\sigma_{m+1} + \hat\sigma_{m+2} + \hat\sigma_{m+3} & > s_{m+3} \\ \quad\ldots \end{cases}$$

Поскольку $s_{m+1} - s_m \in \{0,1\}$, несложно видеть, что первые два условия совместны только если $\hat n_a = s_m$, то есть последняя система эквивалентна следующей:

$$\begin{cases} \hat n_a & = s_m \\ \hat\sigma_{m+1} & > s_{m+1} - s_m \\ \hat\sigma_{m+1} + \hat\sigma_{m+2} & > s_{m+2} - s_m \\ \hat\sigma_{m+1} + \hat\sigma_{m+2} + \hat\sigma_{m+3} & > s_{m+3} - s_m \\ \quad\ldots & \\ \hat\sigma_{m+1} + \cdots + \hat\sigma_{m+K-1} + \hat\sigma_{m+K} & > s_{m+K} - s_m \end{cases} \tag{4.6}$$

Переходя от последней системы к производящему полиному (Утверждение 4.3.6), имеем:

$$\begin{aligned} &\mathbf{P}\big[\,\mathrm{U}_a\bar{\mathrm{U}}_{b_1}\ldots\bar{\mathrm{U}}_{b_K}\big] = \\ &= \big[\,z^\ell a^{s_m} b_1^{>s_{m+1}-s_m}\ldots b_K^{>s_{m+K}-s_m}\big](1+za)^m(1+zb_1\ldots b_K)\ldots(1+zb_K)(1+z)^{L-m-K} = \\ &= \big[\,z^{s_m}a^{s_m}\big](1+za)^m \cdot \big[\,z^{\ell-s_m} b_1^{>s_{m+1}-s_m}\ldots b_K^{>s_{m+K}-s_m}\big] \times \\ &\qquad\qquad \times (1+zb_1\ldots b_K)\ldots(1+zb_K)(1+z)^{L-m-K} = P_1\cdot P_2. \end{aligned}$$

Очевидно,

$$P_1 = \big[\,z^{s_m}a^{s_m}\big](1+za)^m = \tbinom{m}{s_m}.$$

Рассмотрим вторую часть выражения:

$$P_2 = \big[\,z^{\ell-s_m} b_1^{>s_{m+1}-s_m}\ldots b_K^{>s_{m+K}-s_m}\big](1+zb_1\ldots b_K)\ldots(1+zb_K)(1+z)^{L-m-K}. \tag{4.7}$$

Для краткости определим

$$\boldsymbol{u} = \Big((s_{m+1}-s_m),\ \ldots,\ (s_{m+K}-s_m),\ \ldots,\ (s_{m+K}-s_m)\Big)$$

вектор длины $L-m$ и дополним (4.7) фиктивными условиями с $K+1$ по $L-m$. Имеем:

$$P_2 \;=\; \left[\, z^{\ell-s_m} b_1^{>u_1} \dots b_K^{>u_K} b_{K+1}^{>u_{K+1}} \dots b_{L-m}^{>u_{L-m}} \right] \cdot \\ \cdot (1+zb_1 \dots b_{L-m}) \dots (1+zb_{L-m-1}b_{L-m})(1+zb_{L-m}) \;\stackrel{\text{def}}{=}\; \binom{L-m}{\ell-s_m}_{\boldsymbol{u}}.$$

$\binom{L-m}{\ell-s_m}_{\boldsymbol{u}}$ есть число способов выбрать $\ell - s_m$ объектов из $L-m$ (упорядоченных) объектов так, чтобы из первых i объектов было выбрано более u_i. Рекуррентное соотношение для $\binom{L-m}{\ell-s_m}_{\boldsymbol{u}}$ получается аналогично рекуррентному соотношению для биномиальных коэффициентов.

Для начала, отметим, что если $\ell - s_m \leqslant u_{L-m}$, то первое условие $\left[\, z^{\ell-s_m} \right]$ противоречит последнему условию $\left[\, b_{L-m}^{>u_{L-m}} \right]$, и тогда $\binom{L-m}{\ell-s_m}_{\boldsymbol{u}} = 0$. Далее, раскроем последнюю скобку в (4.7). В получившихся двух слагаемых имеем, соответственно, условия $\left[\, \dots b_{L-m-1}^{>u_{L-m-1}} b_{L-m}^{>u_{L-m}} \right]$ и $\left[\, \dots b_{L-m-1}^{>u_{L-m-1}} b_{L-m}^{>u_{L-m}-1} \right]$. Поскольку b_{L-m} и b_{L-m-1} входят в остальные скобки одинаково и $\forall i,\ u_{i+1} - u_i \in \{0,1\}$, то в обоих случаях последнее условие является следствием первого и может быть отброшено. Замечая, что тогда по определению первое слагаемое представляет собой $\binom{L-m-1}{\ell-s_m}_{\boldsymbol{u}}$, а второе $\binom{L-m-1}{\ell-s_m-1}_{\boldsymbol{u}}$, имеем

$$\binom{L-m}{\ell-s_m}_{\boldsymbol{u}} = \left(\binom{L-m-1}{\ell-s_m}_{\boldsymbol{u}} + \binom{L-m-1}{\ell-s_m-1}_{\boldsymbol{u}} \right) \left[\, \ell - s_m > u_{L-m} \right].$$

Повторяя процедуру для каждого из слагаемых имеем рекуррентное соотношение. Граничные условия следуют аналогично граничным условиям для биномиальных коэффициентов.

Итого, число выборок, удовлетворяющих (4.6) есть $P_1 \cdot P_2 = \binom{m}{s_m} \binom{L-m}{\ell-s_m}_{\boldsymbol{u}}$. Лемма доказана. ■

Из рекуррентного соотношения очевидно, что $\binom{L-m}{\ell-s_m}_{\boldsymbol{u}} < \binom{L-m}{\ell-s_m}$. Соответственно, оценка последней леммы лучше оценки

$$h_m(s_m) = \binom{m}{s_m} \binom{L-m}{\ell-s_m} / \binom{L}{\ell},$$

полученной для вклада отдельного алгоритма в Теореме 4.4.3. Если учет в Теореме 4.4.3 одной связи для каждого алгоритма привел к замене «хвоста» гипергеометрического распределения η на гипергеометрическую вероятность $\tilde{\eta}$ в точке , то учет одной цепи, исходящей из алгоритма, приводит в последней лемме к дальнейшему «усечению» этой вероятности.

Для сравнения с оценкой предыдущего параграфа необходимо применить последнюю лемму ко всем алгоритмам в A. Для простоты, сделаем оптимистичное

предположение, что для каждого $a \in A$ среди исходящих из него монотонных цепей, найдется цепь ведущая к верхнему слою A: A_M, $M = \max\limits_{a\in A} n_a$. Назовем такую цепь *максимальной* монотонной цепью. Тогда имеем следующую оценку.

Теорема 4.6.3. *Для любого семейства $\mathcal{F}$, полной выборки $\mathbb{X}$, $|\mathbb{X}| = L$, индикатора ошибки* $\mathrm{I}\colon \mathcal{F} \times \mathbb{X} \to \{0,1\}$, *если в множестве $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ для любого алгоритма a можно найти максимальную цепь, то для любого метода обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$ вероятность переобучения оценивается как:*

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta)] \;\leqslant\; P_{chains}(\eta) \;\stackrel{\text{def}}{=}\; N_0\eta \;+\; \tilde{\eta}\sum_{m=0}^{M} \Delta_m \beta_m,$$

где N_0 — число алгоритмов в последнем слое A_M, $M = \max\limits_{a\in A} n_a$; $\tilde{\eta} = \max\limits_{m} h_m(s_m)$; $\beta_m = \binom{L-m}{\ell-s_m}_{\boldsymbol{u}} \Big/ \binom{L-m}{\ell-s_m} < 1$ и $\binom{L-m}{\ell-s_m}_{\boldsymbol{u}}$ определяется Леммой 4.6.2 при $K = M - m$. Также верна доверительная оценка

$$\mathbf{P}[\, n_{\hat{a}} \geqslant \bar{n}\,(\hat{n}_{\hat{a}}, \eta_{chains}(\alpha))] \;\leqslant\; \alpha,$$

где $\eta_{chains}(\alpha) = \max\{\eta\colon P_{chains}(\eta) \leqslant \alpha\}$.

Доказательство. Для любого $a \in A$ пусть $A_c(a) \subset A$ есть максимальная монотонная цепь из a. Упорядочим алгоритмы в A по убыванию полного числа ошибок, тогда из цепного разложения (4.3) имеем:

$$\mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] \;=\; \sum_{i=1}^{|A|} \mathbf{P}\big[\,\mathrm{U}_{a_i}\bar{\mathrm{U}}_{a_{i-1}}\ldots\bar{\mathrm{U}}_1\big] \;\leqslant\; \sum_{i=1}^{|A|} \mathbf{P}\Big[\mathrm{U}_{a_i} \bigwedge_{a\in A_c(a_i)} \bar{\mathrm{U}}_a\Big]. \tag{4.8}$$

Рассмотрим i-ый член последней суммы, обозначим для краткости $n_{a_i} = m$.

Если $m < M$, то i-ое слагаемое оценивается сверху Леммой 4.6.2, с $K = |A_c(a_i)| = M - m$. Если $m = M$, то имеем оценку $\mathbf{P}[\,\mathrm{U}_{a_i}] \leqslant \eta$. Подставляя оценки в последнее выражение и пользуясь определением профиля расслоения Δ_m, имеем

$$\mathbf{P}\Big[\bigvee_{a\in A} \mathrm{U}_a\Big] \;\leqslant\; N_0\eta + \sum_{m=0}^{M} \Delta_m \frac{\binom{m}{s_m}\binom{L-m}{\ell-s_m}_{\boldsymbol{u}}}{\binom{L}{\ell}}.$$

Пользуясь определением гипергеометрической вероятности $h_m(s_m) = \binom{m}{s_m}\binom{L-m}{\ell-s_m}/\binom{L}{m}$ имеем утверждение теоремы. ■

Отметим, что аналогичная оценка выполняется и для критерия переобучения $\nu - \hat{\nu} \geqslant \varepsilon$ с заменой в теореме $\bar{n}\,(\hat{n}_{\hat{a}}, \eta) \;\to\; L(\hat{n}_{\hat{a}}/\ell + \varepsilon)$ и $\eta \;\to\; P_1(\varepsilon)$.

Пример 4.6.4. Рассмотрим семейство $\mathcal{F}$ с $\mathrm{VCdim}(\mathrm{I}(\mathcal{F})) > L$; тогда A есть полный булев куб, $|A| = 2^L$. Рассмотрим критерий переобучения $\nu_a - \hat{\nu}_a \geqslant \varepsilon$, $\varepsilon = 1\%$, тогда $s_m = \lfloor \frac{\ell}{L} m - \varepsilon\ell \rfloor$. VC-оценка есть

$$\mathbf{P}[\,\mathrm{U}_{\hat{a}}] \;\leqslant\; P_1(\varepsilon) \cdot 2^L.$$

Профиль расслоения для булева куба $D_m = C_L^m$. Оценка последней теоремы есть

$$\mathbf{P}[\,\mathrm{U}_{\hat{a}}] \;\leqslant\; P_1(\varepsilon) \;+\; \tilde{\eta} \sum_{m=0}^{M} C_L^m \beta_m.$$

Численное сравнение с оценкой Теоремы 4.4.3, учитывающей наличие одной связи у каждого алгоритма для разных $L \in (100,\, 1000)$ показывает, что оценка последней теоремы меньше в $3 \div 10$ раз.

Как показывает пример булева куба и эксперимент с семейством линейных классификаторов в главе 6, учет одной монотонной цепи исходящей из каждого алгоритма, вне зависимости от ее длины, не уменьшает оценку существенно в сравнении с учетом одной связи алгоритма.

Можно предположить, что для существенного улучшения оценки необходим учет сходства алгоритмов не столько «в глубину» — крайним случаем которого является максимальная цепь, сколько «в ширину» — крайним случаем которого является единичная окрестность. То есть, учет окрестностей радиуса $\rho > 1$ в A и, в пределе, учет сходства каждого алгоритма со всеми алгоритмами, в которые из него идут монотонные цепи. Интуитивно представляется, что число таких алгоритмов должно расти экспоненциально с ростом размерности пространства параметров $\mathcal{F}$ (а не линейно, как число ближайших соседей $\rho_+(a)$), что может привести к существенному улучшению оценки.

4.7 Выводы

В настоящей главе развивается новый метод учета метрических свойств семейства функций $\mathcal{F}$ в оценках обобщающей способности методов обучения $\mu\colon [\mathbb{X}]^\ell \mapsto \mathcal{F}$. Предлагаются более точные оценки вероятности дизъюнкции событий вида «алгоритм переобучен», чем традиционно используемое в оценках обобщающей способности неравенство Буля, основанные на метрических характеристиках множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ и принципе включения-исключения.

Очевидно, что множества событий рассматриваемые в PAC-оценках обобщающей способности зачастую состоят из существенно пересекающихся друг с другом событий и учет этого пересечения может вести к существенному улучшению оценок. В частности, учет числа «соседей» алгоритма в множестве A в параграфе 4.5 позволяет уменьшить оценку функционала равномерного уклонения частот приблизительно на экспоненциальный множитель.

Предложенный метод приводит к новой комбинаторной характеристике множества алгоритмов A — профилю расслоения-связности $\Delta_{m,q}$ — представляющему

число алгоритмов в A с m ошибками и q (Хэмминговыми) связями с алгоритмами с $m+1$ ошибкой. Профиль расслоения $\Delta_m = \sum_q \Delta_{m,q}$ представляет комбинаторный аналог «распределения вероятности ошибки» (true error distribution), использованного в shell-оценках [47]; профиль связности $\Delta_q^+ = \sum_m \Delta_{m,q}$ не рассматривался ранее в контексте теории обучения и оценок обобщающей способности.

Конечная цель настоящего направления исследований — полный учет структуры множества алгоритмов A для получения точного значения функционала равномерного уклонения. Таким образом, мы фокусируемся на той части завышенности оценок обобщающей способности, которая возникает из-за пренебрежения свойствами семейства $\mathcal{F}$ и может быть устранена без учета конкретного метода обучения.

Возникающий при этом общий вопрос — какие еще комбинаторные параметры множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ помимо уже известных в теории обучения могут играть роль в оценках обобщающей способности. К примеру, в настоящей работе мы не учитываем сходство сразу нескольких алгоритмов между собой в наборах $(a_1, \ldots, a_n) \in A^n$ при $n > 2$ (кроме вырожденного случая цепей алгоритмов). Учет сходства подобного рода соответствует рассмотрению членов высокого порядка в разложении функционала равномерного уклонения Q_A по принципу включения-исключения (4.2).

Для получения более точных оценок Q_A в рамках комбинаторного подхода, безусловно необходимо рассматривать комбинаторные свойства множеств A индуцированных конкретными семействами $\mathcal{F}$. В частности, представляет интерес получение профилей $\Delta_{m,q}$ для широко используемых семейств $\mathcal{F}$ конечной размерности Вапника-Червоненкиса. В более простом случае, получение среднего значения профиля и определение степени его концентрации позволит получить пессимистические оценки профиля подходящие для использования в оценках обобщающей способности.

Другой возможный путь получения оценок профилей — рассмотрение связи между профилем $\Delta_{m,q}$ множества $\mathrm{I}(\mathcal{F}, \mathbb{X})$ и аналогичным профилем множества $\mathrm{I}(\mathcal{F}, X)$ индуцированного на *наблюдаемой* обучающей выборке. Поскольку последний профиль наблюдаем, подобная связь позволила бы получить пессимистичные оценки $\Delta_{m,q}$ на основе обучающей выборки. Оценка такого рода используется для профиля Δ_m в параграфе 3.4.

Глава 5

Характеристики связности семейства линейных классификаторов

Рассмотрим задачу классификации на два класса, $Y = \{-1, +1\}$. Пусть имеется некоторая полная выборка $\mathbb{X} \subset \mathbb{R}^p$. Пусть $\mathbb{X}$ находится в общем положении (для выборок $\mathbb{X}$, взятых из достаточно регулярного распределения в $\mathbb{R}^p$, это выполняется с вероятностью 1). Пополним векторы $x \in \mathbb{X}$ единицами: $x \mapsto (x, 1)$. Будем обозначать через $\langle w, x\rangle$ скалярное произведение пополненного вектора $x \in \mathbb{X}$ и вектора весов $w \in \mathbb{R}^{p+1}$. Для краткости, будем обозначать гиперплоскость $\langle w, x\rangle = 0$ в $\mathbb{R}^{p+1}$ через w.

Рассмотрим семейство линейных классификаторов:

$$\mathcal{F} = \left\{ a_w(x) \stackrel{\text{def}}{=} \operatorname{sign}\langle w, x\rangle \colon w \in \mathbb{R}^{p+1} \right\}. \tag{5.1}$$

Будем далее обозначать a_w функцию – элемент семейства $\mathcal{F}$. Отметим, что компоненты вектора $x \in \mathbb{X}$: $x_j, j = 1, \ldots, p$ могут быть некоторыми нелинейными функциями $x_j = x_j(r)$ некоторого исходного признакового описания r объекта x, поэтому семейство (5.1) представляет в общем случае достаточно широкий класс разделяющих поверхностей.

Пусть целевая функция есть $y \colon \mathbb{X} \to \{-1, +1\}$, тогда индикатор ошибки $I \colon \mathcal{F} \times \times \mathbb{X} \to \{0, 1\}$ естественным образом определяется как

$$I(a_w, x) = [\, a_w(x) \neq y(x)]\,. \tag{5.2}$$

В главе 4 были получены оценки обобщающей способности, точность которых увеличивается с ростом степени связности алгоритмов в A. В связи с этим возникает вопрос о нижних оценках полустепени связности вида

$$\forall a \in A \quad \rho_+(a) = |\mathbb{X}_+(a)| \;\geqslant\; \rho_{\min}.$$

Такие оценки можно было бы, к примеру, использовать вместо профиля связности в Теореме 4.5.7.

Геометрически, связующие объекты $\mathbb{X}_{\pm}(a)$ для линейного классификатора a_w есть те объекты полной выборки $\mathbb{X}$, которых может коснуться разделяющая гиперплоскость w при различных ее «шевелениях», при которых не меняются ответы, которые дает a_w на $\mathbb{X}$. Геометрически также очевидна нижняя оценка для полного числа связей (т.е. общего числа связующих объектов) линейного классификатора

$$\rho_+(a) + \rho_-(a) = |\mathbb{X}_+(a)| + |\mathbb{X}_-(a)| \ > \ p,$$

при $L > p$ (см., к примеру, [14]). Однако нижняя оценка для $\rho_+(a)$ есть тривиально 0. Действительно, для примера можно рассмотреть линейный классификатор $a_w(x)$, дающий неправильные ответы на всех своих связующих объектах.

Более мягкой альтернативой нижней оценке для $\rho_+(a)$ может быть оценка степени концентрации профиля Δ_q^+ возле его среднего значения:

$$|A|^{-1} \sum_{q=0}^{L} [\, q < \bar{q} - \varepsilon] \, \Delta_q^+ \ \leqslant \ \alpha(\varepsilon), \tag{5.3}$$

где $\bar{q} = |A|^{-1} \sum_{q=0}^{L} q\Delta_q^+$ — средняя полустепень связности в A, ε — некоторый небольшой в сравнении с $\bar{q}$ порог и $\alpha(\varepsilon)$ — оценка доли алгоритмов, имеющих полустепень связности ниже $\bar{q} - \varepsilon$.

В настоящей главе выводится среднее значение и оценка дисперсии полустепени связности ρ_+ в семействе линейных классификаторов. Полученные оценки не зависят от полной выборки $\mathbb{X}$ и представляют внутренние комбинаторные характеристики семейства. Для получения оценок используется геометрическая теория конфигураций гиперплоскостей.

5.1 Конфигурации гиперплоскостей

Удобным инструментом для анализа структуры множества $\mathrm{I}(\mathcal{F}, \mathbb{X})$, индуцированного семейством вида $\mathcal{F} = \{\mathrm{sign}(g) \colon g \in \mathcal{G}\}$, где $\mathcal{G}$ есть какое-то достаточно регулярное семейство вещественнозначных функций, является теория геометрических конфигураций. В частности, для линейных классификаторов — конфигураций гиперплоскостей. Общий обзор теории конфигураций можно найти в [12], [32], прекрасное введение в современную теорию конфигураций гиперплоскостей дано в лекциях [62], глубокое алгебраическое изложение представлено в [56], более геометрическое изложение можно найти в более ранних книгах [26], [30].

Рассмотрим семейство $\mathcal{F}$ линейных классификаторов (5.1) с индикатором ошибки (5.2). Обозначим $\mathbb{W} \equiv \mathbb{R}^{p+1}$ пространство параметров семейства $\mathcal{F}$, обозначим для краткости $p+1=d$. Пусть $\mathbb{X} \subset \mathbb{R}^p$ в общем положении.

Определение 5.1.1. *d-мерной конфигурацией однородных гиперплоскостей $\mathcal{H}(\mathbb{X})$ будем называть множество L однородных (проходящих через 0) гиперплоскостей в $\mathbb{W}$:*

$$\mathcal{H}(\mathbb{X}) = \{h_i \colon x_i \in \mathbb{X}\}, \quad h_i = \{w \in \mathbb{W} \colon \langle w, x_i \rangle = 0\}.$$

Будем обозначать $x(h)$ и $h(x)$ взаимно однозначное соответствие между объектами $x \in \mathbb{X}$ и гиперплоскостями $h \in \mathcal{H}(\mathbb{X})$.

Будем далее для краткости писать просто $\mathcal{H}$. Пример конфигурации в $\mathbb{W} = \mathbb{R}^3$ приведен на Рис. 5.1.

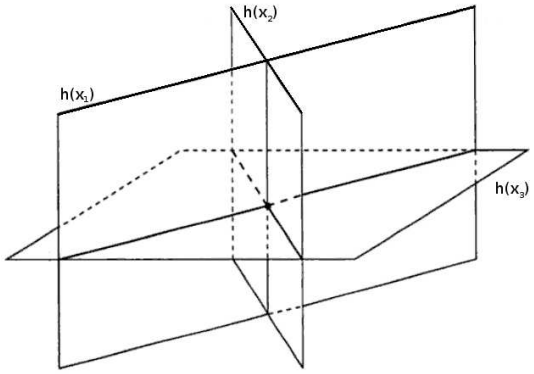

Рис. 5.1. Пример конфигурации 3 плоскостей в $\mathbb{W} = \mathbb{R}^3$, порожденной полной выборкой $\mathbb{X} = \{x_1, x_2, x_3\} \subset \mathbb{R}^2$; $|\mathcal{C}| = 8$ ячеек, $|\mathcal{F}| = 12$ граней, $|\mathcal{E}| = 6$ ребер, $2^3 = 8$ возможных ориентаций плоскостей, т.е. разметок полной выборки.

Отметим, если $\mathbb{X} \in \mathbb{R}^p$ в общем положении, то $\mathcal{H}$ также «в общем положении», то есть, пересечение любых $k \leqslant p$ гиперплоскостей из $\mathcal{H}$ есть линейное подпространство размерности $d-k$ и любые $k > p$ гиперплоскостей пересекаются только в 0.

Утверждение 5.1.2. *Если $\mathbb{X} \subset \mathbb{R}^p$ находится в общем положении, то*

$$\forall \{h_1, \ldots, h_k\} \subset \mathcal{H}(\mathbb{X}), \quad \begin{cases} \dim(h_1 \cap \cdots \cap h_k) = p+1-k, & \text{если } k \leqslant p, \\ h_1 \cap \cdots \cap h_k = \{0\}, & \text{если } k > p, \end{cases}$$

и $h_1 \cap \cdots \cap h_k$ есть линейное подпространство.

Доказательство. $\mathbb{X} \subset \mathbb{R}^p$ в общем положении означает, что любые $k \leqslant p$ векторов $\{x_1, \ldots, x_k\} \subset \mathbb{X}$ линейно независимы и любые $k > p$ векторов линейно зависимы. Следовательно, $p \times k$ матрица $\mathbf{X} = [x_1 \ldots x_k]$ имеет ранг $\min(k, p)$. Тогда $(p+1) \times k$

матрица $\tilde{\mathbf{X}}$, составленная из векторов x_i дополненных единицами единицами, имеет ранг $\min(k, p+1)$ (при $k \leqslant p$ это очевидно; при $k > p$ заметим, что общее положение $\mathbb{X}$ также означает, что никакие $k > p$ точек из $\mathbb{X}$ не принадлежат одной гиперплоскости в $\mathbb{R}^p$). Тогда имеем

$$h_1 \cap \cdots \cap h_k \;\equiv\; \{w\colon \langle w, x_1\rangle = 0, \ldots, \langle w, x_k\rangle = 0\} \;\equiv\; \left\{w\colon w\tilde{\mathbf{X}} = 0\right\}.$$

Множество решений однородной системы $w\tilde{\mathbf{X}} = 0$ из k уравнений с $p+1$ неизвестными есть линейное подпространство размерности $p+1-\operatorname{rank}(\tilde{\mathbf{X}})$ при $\operatorname{rank}(\tilde{\mathbf{X}}) \leqslant p$ и система имеет единственное решение $\{0\}$, при $\operatorname{rank}(\tilde{\mathbf{X}}) \geqslant p+1$. Утверждение доказано. ■

Гиперплоскость h_i разделяет $\mathbb{W}$ на два полупространства, соответствующие классификаторам, дающим правильный/неправильный ответ на объекте x_i.

Определение 5.1.3. *Будем называть, положительным и отрицательным полупространством соответствующим h_i:*

$$h_i^+ \stackrel{\text{def}}{=} \{w\colon \mathrm{I}(a_w, x_i) = 0\} \cup h_i, \qquad h_i^- \stackrel{\text{def}}{=} \{w\colon \mathrm{I}(a_w, x_i) = 1\} \cup h_i.$$

Отметим, что каждое полупространство включает гиперплоскость h_i.

Гиперплоскости $\mathcal{H}$ разбивают пространство $\mathbb{W}$ на множество *ячеек*.

Определение 5.1.4. *Ячейка конфигурации $\mathcal{H}$ есть топологическое замыкание связной компоненты множества $\mathbb{W} \setminus \cup_i h_i$.*

Каждая ячейка представляет d-мерный многогранный бесконечный конус с вершиной в 0. Отметим, что ячейка включает в себя все свои грани всех размерностей.

Утверждение 5.1.5. *Каждой ячейке c взаимно однозначно соответствует алгоритм в $a(c) \in A$:*

$$a(c) = \left(\mathrm{I}(a_w, x_1) \ldots \mathrm{I}(a_w, x_L)\right),$$

где w — любая внутренняя точка ячейки c. Обратно, ячейка, соответствующая алгоритму $a \in A$, есть:

$$c(a) = \operatorname{cl}\left\{w \in \mathbb{W} \,\middle|\, \forall x \in \mathbb{X}\colon \mathrm{I}(a_w, x) = \mathrm{I}(a, x)\right\},$$

где cl *топологическое замыкание.*

Обозначим $\mathcal{C} = \{c(a)\colon a \in A\}$ множество всех ячеек конфигурации $\mathcal{H}$. *Гранями* конфигурации будем называть $(d-1)$-мерные грани многогранников $c \in \mathcal{C}$.

Определение 5.1.6. *Множество граней $\mathcal{F}(h)$ конфигурации $\mathcal{H}$, лежащих в одной гиперплоскости $h \in \mathcal{H}$, есть множество топологических замыканий связных компонент множества $h \setminus \cup_{h_i \neq h} h_i$. Множество всех граней конфигурации $\mathcal{H}$ есть $\mathcal{F} = \cup_{h \in \mathcal{H}} \mathcal{F}(h)$. Множество граней ячейки $c \in \mathcal{C}$ есть $\mathcal{F}(c) = \{f \in \mathcal{F}\colon f \subset c\}$.*

Каждой грани $f \in \mathcal{F}$ можно однозначно поставить в соответствие плоскость $h(f) = h \in \mathcal{H}\colon f \subset h$ в которой лежит грань и соответственно объект $x(f) \in \mathbb{X}$.

Определение 5.1.7. *Будем называть грань $f \in \mathcal{F}(c)$ положительной гранью ячейки c, если ячейка c лежит в положительном полупространстве $h^+(f)$; то есть, алгоритм $a(c)$ дает правильный ответ на объекте $x(f)$.*

Утверждение 5.1.8. *Две смежные (имеющие общую грань f) ячейки c_1, c_2 взаимно однозначно соответствуют двум алгоритмам $a(c_1), a(c_2)$ смежным в графе 1-сходства G^1_A и отличающимся на объекте $x(f)$. Далее 1) каждой грани $f \in \mathcal{F}(c)$ ячейки c взаимно однозначно соответствует связь алгоритма $a(c)$ в графе G^1_A через объект $x(f)$ и, 2) множество положительных граней ячейки c, обозначим его как $\mathcal{F}^+(c)$, соответствует множеству верхних связей алгоритма $a(c)$ в G^1_A, то есть $|\mathcal{F}^+(c)| = \rho_+(a(c))$.*

Ребрами конфигурации $\mathcal{H}$ будем называть $(d-2)$-мерные грани многогранников $c \in \mathcal{C}$.

Определение 5.1.9. *Множество ребер $\mathcal{E}(h, h')$, лежащих на пересечении двух заданных плоскостей $h, h' \in \mathcal{H}, h \neq h'$, есть множество топологически замкнутых связных компонент множества $\{h \cap h'\} \setminus \cup_{h_i \notin \{h,h'\}} h_i$. Множество ребер, лежащих в заданной плоскости h, обозначим как $\mathcal{E}(h) = \cup_{h' \neq h} \mathcal{E}(h, h')$. Множество всех ребер конфигурации есть $\mathcal{E} = \cup_{h,h'} \mathcal{E}(h, h')$.*

Известно, что число ячеек и граней (любой размерности) в d-мерной конфигурации $\mathcal{H}$ для $\mathbb{X}$ в общем положении не зависит от $\mathbb{X}$. Обозначим $C_k(L, d)$ число граней размерности $d-k$ в d-мерной конфигурации из L однородных гиперплоскостей. Тогда число ячеек в конфигурации $\mathcal{H}$ есть [13]:

$$|\mathcal{C}| = C_0(L, d) \stackrel{\text{def}}{=} 2\sum_{k=0}^{d-1} \tbinom{L-1}{k}, \tag{5.4}$$

и число граней есть

$$|\mathcal{F}| = C_1(L, d) = L \cdot C_0(L-1, d-1) = 2L \sum_{k=0}^{d-2} \tbinom{L-2}{k}. \tag{5.5}$$

Последнее нетрудно получить, заметив, что грани, лежащие в каждой из L плоскостей $h \in \mathcal{H}$, в свою очередь являются ячейками $(d-1)$-мерной конфигурации, индуцированной в h пересечениями h и $L-1$ плоскостей $\mathcal{H} \setminus h$.

Отметим, что число ячеек/граней в конфигурации, очевидно, также не зависит от *ориентации* гиперплоскостей, определяемой индикатором ошибки I (и, в конечном счёте, целевой функцией $y\colon \mathbb{X} \to \{-1, +1\}$). При этом профиль расслоения-связности определяется через полустепень связности ρ_+, и чтобы получить для него

оценки, необходимо различать положительную/отрицательную ориентацию граней ячеек. Соответственно, далее будем рассматривать *конфигурации ориентированных гиперплоскостей* (обощением этого понятия являются *ориентированные матроиды* [16]).

Неформально говоря, значение профиля связности Δ_q^+ равно числу ячеек в $\mathcal{H}$, имеющих ровно q положительных граней, а значение профиля расслоения Δ_m равно числу ячеек, лежащих ровно в m отрицательных полупространствах. Последняя величина исследовалась в [51], где для неё была получена верхняя оценка. В настоящей главе рассматривается профиль связности Δ_q^+.

5.2 Средняя связность

Для начала отметим очевидные свойства профилей Δ_q^+, Δ_m, верные для любого семейства $\mathcal{F}$, содержащего наряду с каждой разделяющей поверхностью также и её «инверсию».

Утверждение 5.2.1. *Для семейства линейных классификаторов* (5.1) *с индикатором ошибки* (5.2) *и* $\mathbb{X} \subset \mathbb{R}^p$ *в общем положении 1) профили* Δ_q^+ *и* Δ_q^- *совпадают и 2) профиль* Δ_m *симметричен относительно* $\frac{L}{2}$.

Доказательство. Заметим, что для каждого классификатора $a_w \in \mathcal{F}$ существует его «инверсия» $a_{-w} \in \mathcal{F}$, обозначим $a, \bar{a} \in A$ соответствующие вектора ошибок. Очевидно, что

$$\rho_+(a_w) = \rho_-(a_{-w}), \quad \rho_+(a_{-w}) = \rho_-(a_w), \quad n(a_w, \mathbb{X}) + n(a_{-w}, \mathbb{X}) = L.$$

Тогда

$$\Delta_q^+ \;=\; \sum_{a \in A} [\,\rho_+(a) = q] \;=\; \sum_{a \in A} [\,\rho_-(\bar{a}) = q] \;=\; \sum_{a,a' \in A} [\,\rho_-(a') = q]\,[\,a' = \bar{a}] \;=\; \Delta_q^-$$

и

$$\Delta_m \;=\; \sum_{a \in A} [\,n_a = m] \;=\; \sum_{a \in A} [\,n_{\bar{a}} = L - m] \;=\; \sum_{a,a' \in A} [\,n_{a'} = L - m]\,[\,a' = \bar{a}] \;=\; \Delta_{L-m}.$$

Утверждение доказано. ∎

Далее, несложно получить среднее значение профиля Δ_q^+.

Теорема 5.2.2. *Пусть* $\mathcal{F}$ *есть семейство линейных классификаторов* (5.1) *с индикатором ошибки* (5.2) *и* $\mathbb{X} \subset \mathbb{R}^p$ *в общем положении. Тогда средняя полустепень связности алгоритмов во множестве* $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$

$$\bar{\rho_\pm} = |A|^{-1} \sum_{a \in A} \rho_\pm(a)$$

равна

$$\bar{\rho}_{\pm} = \frac{L \cdot \sum_{k=0}^{p-1} \binom{L-2}{k}}{\sum_{k=0}^{p} \binom{L-1}{k}}.$$

Доказательство. Рассмотрим конфигурацию гиперплоскостей $\mathcal{H}(\mathbb{X})$. Очевидно, суммарное число верхних связей в G_A^1: $\sum_{a\in A} \rho_{\pm}(a)$ равно общему числу ребер G_A^1. Последнее, в свою очередь, равно числу граней $|\mathcal{F}|$ в конфигурации $\mathcal{H}(\mathbb{X})$. Общее число алгоритмов $|A|$ равно числу ячеек $|\mathcal{C}|$ в конфигурации $\mathcal{H}(\mathbb{X})$. Тогда средняя верхняя полустепень связности в A равна $\frac{|\mathcal{F}|}{|\mathcal{C}|}$. То же рассуждение верно для нижней полустепени связности. Используя (5.4), (5.5) имеем утверждение леммы. ■

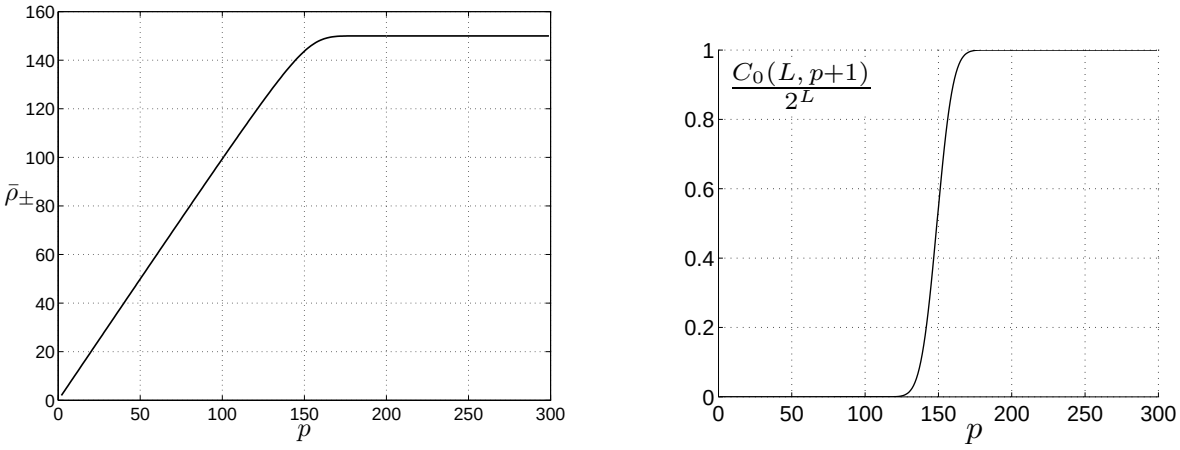

Рис. 5.2. Характеристики множества алгоритмов $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$, индуцированного семейством $\mathcal{F}$ линейных классификаторов на полной выборке $\mathbb{X} \subset \mathbb{R}^p$ размера $|\mathbb{X}| = 300$, в зависимости от размерности p; **Слева:** среднее значение полустепени связности ρ_+ алгоритмов в A; **Справа:** общее число алгоритмов $|A|$ в сравнении с максимально возможным их числом 2^L.

Зависимость $\bar{\rho}_{\pm}$ от размерности пространства параметров $\mathbb{W}$ при фиксированном L приведена слева на Рис.5.2. Средняя полустепень связности растет линейно $\bar{\rho}_{\pm} \approx p$ до значения $p \approx \frac{L}{2}$. Это примерно соответствует уменьшению оценки Теоремы 4.5.7 в 2^{-p} раз в сравнении с VC-оценкой, и согласуется с результатами экспериментального вычисления оценок в главе 6.

«Плато» $p > L/2$ объясняется правым графиком — оно соответствует таким размерностям p, что $A \approx \{0,1\}^L$, то есть семейство $\mathcal{F}$ порождает почти все 2^L возможных классификаций полной выборки $\mathbb{X}$. В этом случае несложно показать, что профиль связности есть $\Delta_q^+ = \binom{L}{L-q}$ и его среднее значение равно $\bar{\rho}_{\pm} = \frac{L}{2}$.

Отметим также, что для произвольного $\mathcal{F}$ существует оценка $\bar{\rho}_{\pm} \leqslant \mathrm{VCdim}(\mathrm{I}(F))$, полученная в [33].

Далее, рассмотрим простейшую характеристику концентрации Δ_q^+ возле среднего значения — дисперсию ρ_+:

$$\mathbf{Var}(\rho_+) \stackrel{\text{def}}{=} \frac{1}{|A|}\sum_{a\in A}\rho_+^2(a) \;-\; \bar{\rho_+}^2$$

Так как $\rho_+(a) = |\mathcal{F}^+(c(a))|$, то

$$\sum_{a\in A}\rho_+^2(a) \;=\; \sum_{c\in\mathcal{C}}|\mathcal{F}^+(c)|^2\,,$$

и задача оценивания дисперсии связности сводится к задаче оценивания суммы квадратов числа положительных граней по ячейкам конфигурации $\mathcal{H}$. Последняя может быть решена при помощи теоремы о зоне гиперплоскости (zone theorem) из теории конфигураций гиперплоскостей.

5.3 Зона гиперплоскости в конфигурации

Пусть $h \in \mathcal{H}$ есть некоторая плоскость конфигурации $\mathcal{H}$ и $x(h) \in \mathbb{X}$ есть соответствующий ей объект полной выборки. Рассмотрим конфигурацию плоскостей $\mathcal{H}' = \mathcal{H}\backslash\{h\}$, пусть $\mathcal{C}', \mathcal{F}'$ её ячейки и грани. Пример такой конфигурации в $\mathbb{R}^2$ приведен на Рис. 5.3. В примере для наглядности использована проективная конфигурация $\mathcal{H}_p$ *неоднородных* линий в $\mathbb{R}^2$, отметим, что каждой такой конфигурации соответствует конфигурация $\mathcal{H}$ *однородных* плоскостей в $\mathbb{R}^3$, если рассматривать конфигурацию $\mathcal{H}_p$ как пересечение конфигурации $\mathcal{H}$ с проективной плоскостью $z = 1$. Пример такого проективного соответствия конфигураций приведен на Рис. 5.4.

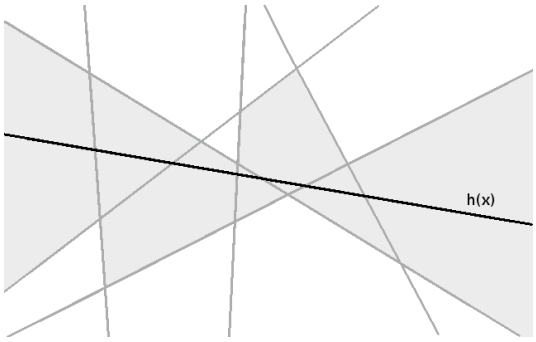

Рис. 5.3. Пример конфигурации $\mathcal{H}\backslash\{h(x)\}$ с удалением одной гиперплоскости и зоны гиперплоскости $h(x)$ в конфигурации $\mathcal{H}\setminus\{h(x)\}$ в $\mathbb{R}^2$ на примере неоднородных гиперплоскостей.

Ячейки $\mathcal{C}'$ соответствуют множеству алгоритмов $A' = \mathrm{I}(\mathcal{F}, \mathbb{X}')$, $\mathbb{X}' = \mathbb{X}\setminus\{x\}$, получаемому при удалении из полной выборки одного объекта и соответствующего ему бита из всех векторов ошибок A. При этом пары алгоритмов в A, отличающихся только на объекте x, дают в A' один алгоритм.

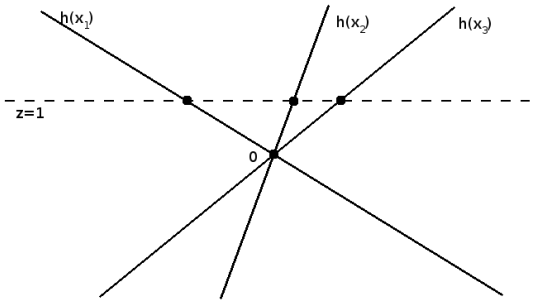

Рис. 5.4. Пример *неоднородной* конфигурации индуцированной в проективной плоскости $\mathbb{R}^1$ $(z = 1)$ *однородной* конфигурацией в $\mathbb{R}^2$.

Утверждение 5.3.1. *Ячейки конфигурации $\mathcal{H}$ получаются пересечением ячеек конфигурации $\mathcal{H}'$ с полупространствами h^+, h^-:*

$$\mathcal{C} = \{c' \cap h^+ \colon c' \in \mathcal{C}'\} + \{c' \cap h^- \colon c' \in \mathcal{C}'\}.$$

Грани конфигурации $\mathcal{H}$ получаются 1) пересечением граней $\mathcal{H}'$ с h^+, h^- и 2) пересечением ячеек $\mathcal{H}'$ с плоскостью h:

$$\mathcal{F} = \{f' \cap h^+ \colon f' \in \mathcal{F}'\} + \{f' \cap h^- \colon f' \in \mathcal{F}'\} + \{f' \cap h \colon f' \in \mathcal{F}'\}.$$

Определение 5.3.2. *Будем называть $c' \cap h^\pm$, $f' \cap h^\pm$ положительным / отрицательным усечением ячейки / грани конфигурации $\mathcal{H}'$.*

Утверждение 5.3.3. *Для любой ячейки $c' \in \mathcal{C}'$ либо оба усечения $c' \cap h^\pm$ непусты (если h пересекает c'), либо одно из них непусто. То же справедливо для граней $f' \in \mathcal{F}'$. Таким образом, усечение ставит в соответствие каждой ячейке/грани $c' \in \mathcal{C}'$, $f' \in \mathcal{F}'$ одну или две (если h пересекает c', f') ячейки/грани в $\mathcal{C}$, $\mathcal{F}$.*

Усечение ставит в соответствие каждому алгоритму $a' \in A'$ один или два алгоритма в A, получающихся добавлением бита, соответствующего объекту x. Аналогично, каждой каждой связи алгоритмов в A' усечение ставит в соответствие одну или две связи в A. Определим обратное соответствие.

Определение 5.3.4. *Каждой ячейке $c \in \mathcal{C}$ можно однозначно поставить в соответствие её «дополнение» в $\mathcal{C}'$:*

$$\tilde{c}(c) \stackrel{\text{def}}{=} \{c' \in \mathcal{C}' \colon c' \cap h^+ = c \vee c' \cap h^- = c\}.$$

Каждой грани $f \in \mathcal{F}$, не лежащей в h, однозначно соответствует её «дополнение» в $\mathcal{F}'$:

$$\tilde{f}(f) \stackrel{\text{def}}{=} \{f' \in \mathcal{F}' \colon f' \cap h^+ = f \vee f' \cap h^- = f\}.$$

Дополнение однозначно ставит в соответствие каждому алгоритму $a \in A$ алгоритм $a'(a) \in A'$, получающийся из a удалением бита, соответствующего объекту x.

Далее, отметим, что

Утверждение 5.3.5. *Пересечения ячеек/граней конфигурации $\mathcal{H}'$ с плоскостью h либо пусты, либо представляют собой, соответственно, грани/рёбра конфигурации $\mathcal{H}$ лежащие в h.*

Определение 5.3.6. *Обозначим*

$$[\, c' \times h] \stackrel{\text{def}}{=} [\, c' \cap h \in \mathcal{F}(h)]$$

индикатор того, что плоскость h пересекает ячейку $c' \in \mathcal{C}'$ и

$$[\, f' \times h] \stackrel{\text{def}}{=} [\, f' \cap h \in \mathcal{E}'(h)]$$

индикатор того, что h пересекает $f' \in \mathcal{F}'$.

Определение 5.3.7. *Обозначим*

$$[\, c | h^{\pm}] \stackrel{\text{def}}{=} [\, c \cap h = \mathcal{F}^{\pm}(c)]$$

индикатор того, что ячейка c смежна с положительной/отрицательной стороной плоскости h.

Смежность ячейки c с плоскостью h означает, что в A есть алгоритм отличающийся от $a(c)$ только на объекте $x(h)$.

Определение 5.3.8. *Множество ячеек конфигурации, которые пересекает заданная гиперплоскость, не принадлежащая конфигурации, называется зоной гиперплоскости в конфигурации. Множество ячеек зоны гиперплоскости h в конфигурации $\mathcal{H}'$ есть*

$$Zone(\mathcal{H}', h) \stackrel{\text{def}}{=} \{c' \in \mathcal{C}' \colon c' \times h\}\,.$$

Множество ячеек $Zone(\mathcal{H}', h)$ соответствует множеству тех алгоритмов в A', которые «расщепятся» на два алгоритма при добавлении в полную выборку объекта x.

Определение 5.3.9. *Сложностью зоны будем называть сумму числа граней по ячейкам зоны*

$$Z_1(\mathcal{H}', h) \;=\; \sum_{c' \in Zone(\mathcal{H}', h)} |\mathcal{F}'(c')|\,.$$

Положительной/отрицательной полусложностью зоны будем называть сумму числа только таких граней, которые частично или полностью лежат, соответственно, в полупространстве h^+ и h^- :

$$Z_1^{\pm}(\mathcal{H}', h) \;=\; \sum_{c' \in Zone(\mathcal{H}', h)} \sum_{f'} [\, f' \in \mathcal{F}'(c')] \, [\, f' \cap h^{\pm} \neq \varnothing] \,. \tag{5.6}$$

Отметим, что грани, $f' \in \mathcal{F}'$, пересекаемые h, учитываются в $Z_1(\mathcal{H}', h)$ и $Z_1^{\pm}(\mathcal{H}', h)$ дважды — по разу в каждой из ячеек которым принадлежит грань.

В терминах множества алгоритмов A, сложность зоны h в $\mathcal{H}'$ есть суммарное число связей тех алгоритмов в A, которые имеют связь через объект $x(h)$, за вычетом самих связей через объект x. Положительная/отрицательная полусложность зоны есть то же, только для алгоритмов, которые которые при этом не ошибаются/ошибаются на x.

Определение 5.3.10. *Обозначим максимально возможную сложность (или полусложность) зоны в конфигурации из $(L-1)$ однородных гиперплоскостей в $\mathbb{R}^d$ как*

$$Z_1^{\pm}(L-1, d) \;\stackrel{\text{def}}{=}\; \max_{\mathbb{X}, x \in \mathbb{X}} Z_1^{\pm}(\mathcal{H}'(\mathbb{X}, x), h(x)).$$

Обозначим $z_1(L-1, d)$ максимальную сложность зоны в конфигурации $(L-1)$ неоднородных гиперплоскостей в общем положении.

Для последней величины имеем следующую оценку, полученную в [27].

Лемма 5.3.11 (Zone theorem [27]).

$$z_1(L-1, d) \;\leqslant\; 4(d-1)\tbinom{L-1}{d-1} + 2\sum_{j=1}^{d-1} j \tbinom{L-1}{j}$$

Из последней леммы можно получить оценку для сложности зоны $Z_1(L-1, d)$ в конфигурации *однородных* гиперплоскостей. Для этого используем связь между однородной d-мерной конфигурацией $\mathcal{H}'$ и соответствующей ей неоднородной $(d-1)$-мерной конфигурацией $\mathcal{H}'_p$ в проективной гиперплоскости.

Лемма 5.3.12. *Максимальное число граней зоны в конфигурации однородных гиперплоскостей и максимальное число граней зоны в конфигурации неоднородных гиперплоскостей связаны как:*

$$Z_1(L-1, d) \;\leqslant\; 2\, z_1(L-2, d-1) + C_0(L-1, d-1).$$

Доказательство. Рассмотрим d-мерную конфигурацию $\mathcal{H}'$ из $(L-1)$ однородных гиперплоскостей и однородную плоскость h пересекающую $\mathcal{H}'$. Выберем одну из гиперплоскостей в $\mathcal{H}'$, обозначим её h'_∞, и проведем произвольную параллельную ей проективную гиперплоскость h'_p.

Тогда в плоскости h'_p имеем конфигурацию из $(L-2)$ неоднородных $(d-1)$-мерных гиперплоскостей $\mathcal{H}'_p = \left\{h'_p \cap h' \colon h' \in \mathcal{H}',\ h' \neq h'_\infty\right\}$, индуцированную пересечениями h'_p с плоскостями $\mathcal{H}'$.

В силу центральной симметрии $\mathcal{H}'$, если элемент e' (ячейка, грань, ребро) принадлежит конфигурации, элемент $-e'$ также принадлежит конфигурации, т.е. элементы $\mathcal{H}'$ можно разбить на пары. За счет выбора проективной плоскости h'_p параллельно одной из плоскостей в $\mathcal{H}'$, имеем, что если e' пересекается с h'_p, то $-e'$ не пересекается с h'_p и обратно.

Между точками в h'_p и однородными прямыми в $\mathbb{W}$ (кроме прямых, лежащих в h'_∞) устанавливается взаимно однозначное проективное соответствие — точке $P \in h'_p$ соответствует прямая $-P0P$. Тогда каждой паре элементов $(e', -e')$ конфигурации $\mathcal{H}'$ (кроме тех пар, которые лежат в плоскости h'_∞) взаимно однозначно соответствует элемент $e'_p = e' \cap h'_p$ конфигурации $\mathcal{H}'_p$. Это соответствие, очевидно, сохраняет отношение принадлежности между элементами: если некоторый элемент e'_p конфигурации $\mathcal{H}'_p$ принадлежит некоторому другому элементу d'_p, то $e' \subset d'$ и $-e' \subset -d'$; обратное также верно.

Далее, заметим, что вследствие центральной симметрии $\mathcal{H}'$, если грань f' конфигурации $\mathcal{H}'$ является гранью зоны гиперплоскости h, то грань $-f'$ также является гранью зоны h. Таким образом, каждой паре $(f', -f')$ граней зоны h в $\mathcal{H}'$ (кроме пар граней, лежащих в h'_∞) взаимно однозначно соответствует грань $f' \cap h'_p$ зоны гиперплоскости h_p в конфигурации $\mathcal{H}'_p$. Имеем, что сложность зоны

$$Z_1(\mathcal{H}', h) = 2z_1(\mathcal{H}'_p, h_p) + \#\left\{\text{граней } Zone(\mathcal{H}', h), \text{ лежащих в } h'_\infty\right\}.$$

Заметим, что число граней $Zone(\mathcal{H}', h)$, лежащих в h'_∞, не превышает общего числа граней конфигурации $\mathcal{H}'$, лежащих в h'_∞, то есть $C_0(L-2, d-1)$. Отсюда имеем

$$Z_1(\mathcal{H}', h) \leqslant 2z_1(\mathcal{H}'_p, h_p) + C_0(L-2, d-1),$$

из чего следует утверждение леммы. ■

5.4 Дисперсия связности

Для оценки дисперсии $\rho_+(a)$ необходимо оценить величину $\sum_{a\in A} \rho_+^2(a)$, равную суммарному числу *пар* положительных связей всех алгоритмов в A.

Идея оценки в том, чтобы для каждого объекта $x \in \mathbb{X}$ рассмотреть подмножество таких алгоритмов, что для каждого из них в A существует алгоритм, отличающийся от него только на x, и оценить суммарное число связей этих алгоритмов при помощи

теоремы о зоне гиперплоскости из предыдущего параграфа. Это число связей в свою очередь равно числу тех из интересующих нас *пар* связей, в которых одна из связей идет через объект x. Суммируя такие оценки по объектам $x \in \mathbb{X}$, получаем требуемое общее число пар связей.

Пусть, как и в предшествующем параграфе, $\mathcal{H}' = \mathcal{H}\backslash\{h\}$ и $\mathcal{C}', \mathcal{F}'$ — ячейки/грани конфигурации $\mathcal{H}'$. Для начала приведём для удобства несколько геометрически очевидных соотношений, которые будут использованы при доказательстве следующей теоремы. Все приведённые соотношения имеют прозрачную интерпретацию в терминах множеств алгоритмов A, A'. Индикаторные функции здесь определены на множествах $\mathcal{F}, \mathcal{C}, \mathcal{H}, \mathcal{F}', \mathcal{C}', \mathcal{H}'$ и их декартовых произведениях.

$$\forall c \in \mathcal{C}, \quad |\mathcal{F}^+(c)| \;=\; \sum_{f\in\mathcal{F}} [\, f \in \mathcal{F}^+(c)] \;=\; \sum_{h\in\mathcal{H}} [\, c|h^+] \tag{5.7}$$

— число положительных граней ячейки равно числу плоскостей, с которыми ячейка смежна с положительной стороны.

$$\forall c \in \mathcal{C}, h \in \mathcal{H}, \quad [\, c|h^+] \;=\; \sum_{f\in\mathcal{F}} [\, f \in \mathcal{F}^+(c)]\, [\, f \in \mathcal{F}(h)] \tag{5.8}$$

— у ячейки c смежной с положительной стороной h существует единственная положительная грань, лежащая в h. Суммируя (5.8) по c имеем:

$$\forall h \in \mathcal{H}, \quad \sum_{c} [\, c|h^+] \;=\; |\mathcal{F}(h)| \tag{5.9}$$

— число ячеек, смежных с положительной стороной h, равно числу граней конфигурации, лежащих в h.

$$\forall f \in \mathcal{F}, h \in \mathcal{H}, \quad [\, f \notin \mathcal{F}(h)] \;=\; \sum_{f'} \Big([\, f' \cap h^+ = f] + [\, f' \cap h^- = f] \Big) \tag{5.10}$$

— для грани f конфигурации $\mathcal{H}$, не лежащей в плоскости h, существует единственная грань f' конфигурации $\mathcal{H}\backslash\{h\}$, из которой f получается либо положительным, либо отрицательным усечением f'.

$$\forall c \in \mathcal{C}, h \in \mathcal{H} \quad [\, c|h^+] \;=\; \sum_{c'} [\, c' \cap h^+ = c]\, [\, c' \times h] \tag{5.11}$$

— для ячейки c конфигурации $\mathcal{H}$, смежной с положительной стороной h, существует единственная ячейка c' конфигурации $\mathcal{H}\backslash\{h\}$, пересекаемая плоскостью h, из которой c получается положительным усечением.

$$\forall h \in \mathcal{H}, c' \in \mathcal{C}', f' \in \mathcal{F}', \quad c' \cap h^{\pm} \in \mathcal{C} \cup \{\varnothing\}\,,\; f' \cap h^{\pm} \in \mathcal{F} \cup \{\varnothing\} \tag{5.12}$$

— усечение ячейки/грани конфигурации $\mathcal{H}'$, либо представляет ячейку/грань конфигурации $\mathcal{H}$, либо пусто.

$$\forall h \in \mathcal{H}, f' \in \mathcal{F}', c' \in \mathcal{C}' \quad [\, f' \cap h^+ \in \mathcal{F}(c' \cap h^+)] \;=\; [\, f' \in \mathcal{F}'(c')]\, [\, f' \cap h^+ \neq \varnothing] \tag{5.13}$$

— «усечение» грани f' принадлежит «усечению» ячейки c', если и только если грань f' принадлежит ячейке c' и усечение грани непусто.

$$\forall f' \in \mathcal{F}', c' \in \mathcal{C}' \quad [\, f' \in \mathcal{F}'(c')] \;=\; [\, f' \in \mathcal{F}'^{h^+}(c')] + [\, f' \in \mathcal{F}'^{h^-}(c')] \tag{5.14}$$

— грань ячейки может быть либо положительной, либо отрицательной.

$$\forall h \in \mathcal{H}, f' \in \mathcal{F}' \quad [\, f' \cap h^+ \neq \varnothing]\, [\, f' \cap h^- \neq \varnothing] \;=\; [\, f' \cap h \in \mathcal{E}(h)] \tag{5.15}$$

— грань f' пересекается с обоими полупространствами h^+, h^-, если и только если грань пересекается с плоскостью h.

$$\forall h \in \mathcal{H}, f' \in \mathcal{F}', f' \times h \quad \sum_{c' \in \mathcal{C}'} [\, c' \times h]\, [\, f' \in \mathcal{F}'^{h^-}(c')] = 1 \tag{5.16}$$

— для грани f', пересекающейся с h, существует единственная ячейка c', пересекающаяся с h, в которой эта грань отрицательна.

Имеем следующую оценку:

Теорема 5.4.1. *Пусть $\mathcal{F}$ есть семейство линейных классификаторов* (5.1) *с индикатором ошибки* (5.2) *и $\mathbb{X} \subset \mathbb{R}^p$ в общем положении. Тогда дисперсия полустепени связности алгоритмов во множестве $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$*

$$\mathbf{Var}_a(\rho_\pm(a)) = |A|^{-1} \sum_{a \in A} (\rho_\pm(a) - \bar{\rho_\pm})^2$$

оценивается сверху как

$$\mathbf{Var}_a(\rho_\pm(a)) \;\leqslant\; L \cdot \frac{C_0(L-1, d-1) + Z_1^+(L-1, d) - C_1(L-1, d-1)}{C_0(L, d)} \;- \frac{C_1(L, d)^2}{C_0(L, d)^2},$$

где Z_1^+ – максимальная полусложность зоны из Определения 5.3.10.

Доказательство. Поскольку профили Δ_q^+ и Δ_q^- совпадают, будем рассматривать только дисперсию $\rho_+(a)$.

Имеем

$$\mathbf{Var}_a(\rho_+(a)) = \frac{1}{|A|} \sum_{a \in A} \rho_+^2(a) \;-\; \bar{\rho_+}^2.$$

Ввиду Теоремы 5.2.2 нам остается оценить только первое слагаемое.

Рассмотрим конфигурацию гиперплоскостей $\mathcal{H}(\mathbb{X})$, пусть $\mathcal{C}, \mathcal{F}$ её ячейки и грани; везде далее суммирование по h, c, f идет по $\mathcal{H}, \mathcal{C}, \mathcal{F}$ соответственно. Имеем $\rho_+(a) = |\mathcal{F}^+(c(a))|$, следовательно:

$$\sum_{a\in A}\rho_+^2(a) = \sum_c |\mathcal{F}^+(c)|\cdot|\mathcal{F}^+(c)| \overset{(5.7)}{=} \sum_c\sum_h [\,c|h^+]\sum_f [\,f\in\mathcal{F}^+(c)] =$$
$$\overset{(5.8)}{=} \sum_h\sum_c [\,c|h^+]\left([\,c|h^+] + \sum_f [\,f\in\mathcal{F}^+(c)]\,[\,f\notin\mathcal{F}(h)]\right) =$$
$$= \sum_h\left(\sum_c [\,c|h^+] + \sum_c [\,c|h^+]\sum_f [\,f\in\mathcal{F}^+(c)]\,[\,f\notin\mathcal{F}(h)]\right) =$$
$$= \sum_h\Big((\mathrm{I}) + (\mathrm{II})\Big)$$

Из (5.9) следует, что первое слагаемое в последней сумме есть число граней, лежащих в плоскости h:

$$(\mathrm{I}) = |\mathcal{F}(h)| = C_0(L-1, d-1).$$

Рассмотрим второе слагаемое:

$$(\mathrm{II}) = \sum_c [\,c|h^+]\sum_f [\,f\in\mathcal{F}^+(c)]\,[\,f\notin\mathcal{F}(h)]\,.$$

(II) представляет собой сумму числа положительных граней, не лежащих в h, у ячеек, смежных с h с положительной стороны. Перейдём от суммирования по ячейкам и граням $\mathcal{H}$ к суммированию по ячейкам и граням конфигурации $\mathcal{H}'$. Имеем:

$$(\mathrm{II}) \overset{(5.10)}{\underset{(5.11)}{=}} \sum_{f,c}\sum_{c'} [\,c'\cap h^+ = c]\,[\,c'\times h]\sum_{f'}\Big([\,f'\cap h^+ = f] + [\,f'\cap h^- = f]\Big)\,[\,f\in\mathcal{F}^+(c)] =$$
$$= \sum_{f',c'} [\,c'\times h]\sum_{f,c} [\,c'\cap h^+ = c]\left[\,f'\cap h^+ = f\right][\,f\in\mathcal{F}^+(c)] =$$
$$\overset{(5.12)}{=} \sum_{f',c'} [\,c'\times h]\,[\,f'\cap h^+\neq\varnothing]\,[\,f'\cap h^+\in\mathcal{F}^+(c'\cap h^+)] =$$
$$\overset{(5.13)}{=} \sum_{c'} [\,c'\times h]\sum_{f'} [\,f'\cap h^+\neq\varnothing]\,[\,f'\in\mathcal{F}'^+(c')] = (\mathrm{II}).$$

Второе равенство здесь следует из того, что $[\,f'\cap h^- = f]\,[\,c'\cap h^+ = c]\,[\,f\in\mathcal{F}^+(c)] = 0$, т.к. условия противоречат друг другу. Заметим, что мы получили сумму по положительным граням полузоны плоскости h в конфигурации $\mathcal{H}'$. Заметим далее, что за счет суммирования только по положительным граням, те грани полузоны, которые пересекаются плоскостью h, учитываются в последней сумме единожды, а

не дважды, как в (5.6). Поскольку число таких граней известно (оно равно числу рёбер из $\mathcal{E}$ лежащих в h), мы можем учесть этот факт в оценке последней суммы. Продолжая цепочку равенств, имеем:

$$\text{(II)} \overset{(5.14)}{=} \sum_{c'} [\, c' \times h] \sum_{f'} [\, f' \cap h^+ \neq \varnothing] \, [\, f' \in \mathcal{F}'(c')] - \\ - \sum_{c'} [\, c' \times h] \sum_{f'} [\, f' \cap h^+ \neq \varnothing] \, [\, f' \in \mathcal{F}'^{-}(c')] = \text{(IIa)} - \text{(IIb)}.$$

Здесь (IIa) по определению есть сложность полузоны $Z_1^+(\mathcal{H}', h)$. Поправка (IIb) есть сумма по отрицательным граням ячеек полузоны, очевидно, она не меньше, чем сумма по тем отрицательным граням полузоны, которые пересекаются с h:

$$\text{(IIb)} \geqslant \sum_{c'} [\, c' \times h] \sum_{f'} [\, f' \cap h^+ \neq \varnothing] \, [\, f' \cap h^+ \neq \varnothing] \, [\, f' \in \mathcal{F}'^{-}(c')] = \\ \overset{(5.15)}{=} \sum_{f'} [\, f' \cap h \in \mathcal{E}(h)] \sum_{c'} [\, c' \times h] \, [\, f' \in \mathcal{F}'^{-}(c')] \overset{(5.16)}{=} \sum_{f'} [\, f' \cap h \in \mathcal{E}(h)] = \\ = |\mathcal{E}(h)| = C_1(L-1, d-1).$$

Итого имеем:

$$\text{(II)} \;\leqslant\; Z_1^+(\mathcal{H}', h) - C_1(L-1, d-1) \;\leqslant\; Z_1^+(L-1, d) - C_1(L-1, d-1)$$

и

$$\sum_{a \in A} \rho_+^2(a) = \sum_h \Big(\text{(I)} + \text{(II)} \Big) \leqslant L \Big(C_0(L-1, d-1) + Z_1^+(L-1, d) - C_1(L-1, d-1) \Big).$$

Используя $|A| = |\mathcal{C}| = C_0(L, d)$ и Теорему 5.2.2 имеем утверждение теоремы. ∎

Верхняя оценка для стандартного уклонения $\sqrt{\mathbf{Var}_a \rho_\pm(a)}$ полустепени связности в A от среднего значения, следующая из последней теоремы, дана пунктирной линией на Рисунке 5.4. Здесь полусложность зоны оценена сверху полной сложностью: $Z_1^+(L-1, d) < Z_1(L-1, d)$.

К сожалению, оценка не достаточно точна, чтобы использовать её для получения оценок концентрации профиля Δ_q^+ вида (5.3). Отметим, однако, что оценка последней теоремы отражает точное значение дисперсии на «плато» $p > L/2$. А именно, для $A = \{0, 1\}^L$ имеем $\Delta_q^+ = \binom{L}{L-q}$ и точное значение $\sqrt{\mathbf{Var}_a \rho_\pm(a)} = \frac{\sqrt{L}}{2}$, что для случая приведенного на Рисунке 5.2 дает $\frac{\sqrt{L}}{2} \approx 8.6$ и совпадает со значением оценки.

Одна из причин завышенности оценки $\mathbf{Var}_a \rho_\pm(a)$ — в оценке полусложности зоны полной сложностью и в грубости перехода от оценки для неоднородной конфигурации к оценке для однородной конфигурации в Лемме 5.3.12. Соответственно, оценка

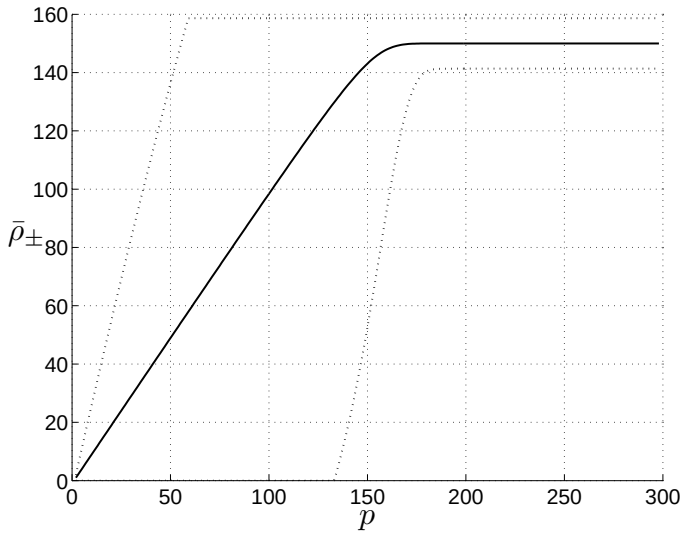

Рис. 5.5. Характеристики множества алгоритмов $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$, индуцированного семейством $\mathcal{F}$ линейных классификаторов на полной выборке $\mathbb{X} \subset \mathbb{R}^p$ размера $|\mathbb{X}| = 300$, в зависимости от размерности p; среднее (сплошная линия) и верхняя оценка стандартного отклонения (пунктирная линия) полустепени связности ρ_+ алгоритмов в A.

для $\mathbf{Var}_a \rho_{\pm}(a)$ может быть существенно улучшена, если непосредственно оценить полусложность зоны в однородной конфигурации аналогично тому, как это сделано для сложности зоны в неоднородной конфигурации в [27]. Предположительно, дисперсия полустепени связности для всех значений p должна быть приблизительно такой же величины, как на «плато».

Отметим, что хотя точная форма профиля Δ_q^+, по-видимому, зависит как от $\mathcal{F}$, так и от $\mathbb{X}$, его среднее значение и оценка дисперсии, полученные в Теоремах 5.2.2, 5.4.1, не зависят от $\mathbb{X}$ и являются внутренними комбинаторными свойствами семейства линейных классификаторов $\mathcal{F}$.

Глава 6

Эксперименты с семейством линейных классификаторов

Рассмотрим задачу классификации на два класса, $Y = \{-1, +1\}$ с полной выборкой $\mathbb{X} \subset \mathbb{R}^d$ в общем положении и семейством линейных классификаторов (5.1), (5.2) введенным в главе 5. Фиксируем также метод обучения $\mu\colon [\mathbb{X}]^\ell \to \mathcal{F}$ — метод опорных векторов (SVM) с линейным ядром.

6.1 Оценки профилей расслоения-связности по методу Монте-Карло

Для вычисления оценок Теорем 3.3.1, 3.3.3, 3.4.3, 4.4.3, 4.5.7, 4.6.3 необходимо вычислить профили

$$\Delta_{m,q} = \operatorname{card}\{a \in A\colon n_a = m,\ \rho_+(a) = q\},$$
$$\Delta_m = \operatorname{card}\{a \in A\colon n_a = m\},$$
$$\hat{\Delta}_s = \operatorname{card}\{a \in A\colon \hat{n}_a = s\},$$

последний — для некоторой случайно выбранной $X \in [\mathbb{X}]^\ell$.

Мощность A для практических значений L слишком велика ($10^{10} \div 10^{30}$), чтобы вычислить профили точно, поэтому мы используем выборку алгоритмов $\tilde{A} \subset A$ для получения несмещённых оценок:

$$\tilde{\Delta}_{m,q} = \tfrac{|A|}{|\tilde{A}|} \cdot \operatorname{card}\big\{\tilde{a} \in \tilde{A}\colon \rho_+(\tilde{a}) = q,\ n_{\tilde{a}} = m\big\},$$
$$\tilde{\Delta}_m = \tfrac{|A|}{|\tilde{A}|} \cdot \operatorname{card}\big\{\tilde{a} \in \tilde{A}\colon n_{\tilde{a}} = m\big\},$$
$$\tilde{\Delta}_s = \tfrac{|A|}{|\tilde{A}|} \cdot \operatorname{card}\big\{\tilde{a} \in \tilde{A}\colon \hat{n}_{\tilde{a}} = s\big\}.$$

Процедура выбора подмножества $\tilde{A} \subset A$ из равномерного распределения на A для случая линейных классификаторов описана в параграфе 6.2. Примеры профилей представлены на Рис. 6.1.

Важное экспериментальное наблюдение здесь состоит в том, что профиль $\tilde{\Delta}_{m,q}$ почти разделим

$$\tilde{\Delta}_{m,q} \approx \tilde{\Delta}_m \, \tilde{\Delta}_q^+ \,/\, |A| \,.$$

Для проверки для нескольких случайно сгенерированных $\mathbb{X}$ были рассчитаны следующие показатели:

1. сумма остатков

$$\sum_{m,q} (\tilde{\Delta}_{m,q} - \tilde{\Delta}_m \, \tilde{\Delta}_q^+ \,/\, |A|),$$

2. сумма квадратов остатков в процентах от суммы квадратов значений $\tilde{\Delta}_{m,q}$

$$\frac{\sum_{m,q} (\tilde{\Delta}_{m,q} - \tilde{\Delta}_m \, \tilde{\Delta}_q^+ \,/\, |A|)^2}{\sum_{m,q} \tilde{\Delta}_{m,q}^2}.$$

Для всех сгенерированных $\mathbb{X}$ первое оказывается равно нулю, второе $\approx 1\%$ (то есть дисперсия остатков близка к нулю). Поскольку $\mathbb{X}$ были взяты случайно, возникает гипотеза:

Гипотеза 6.1.1. Для любой $\mathbb{X}$ в общем положении профиль расслоения-связности семейства линейных классификаторов равен

$$\Delta_{m,q} = \Delta_m \, \Delta_q^+ \,/\, |A| \,.$$

6.2 Процедура сэмплинга алгоритмов из множества A

Трудность заключается в том, что необходим равномерный сэмплинг из множества бинарных векторов A, а не из параметризованного множества гиперплоскостей $\mathcal{F}$. Простейший способ: выбрать равновероятно один из бинарных векторов из $\{0,1\}^L$ и определить, принадлежит ли он множеству A. Однако последний шаг требует решения системы L неравенств, что делает такой способ сэмплинга достаточно медленным для оценок по методу Монте-Карло.

Используем для выбора алгоритмов из A следующую процедуру.

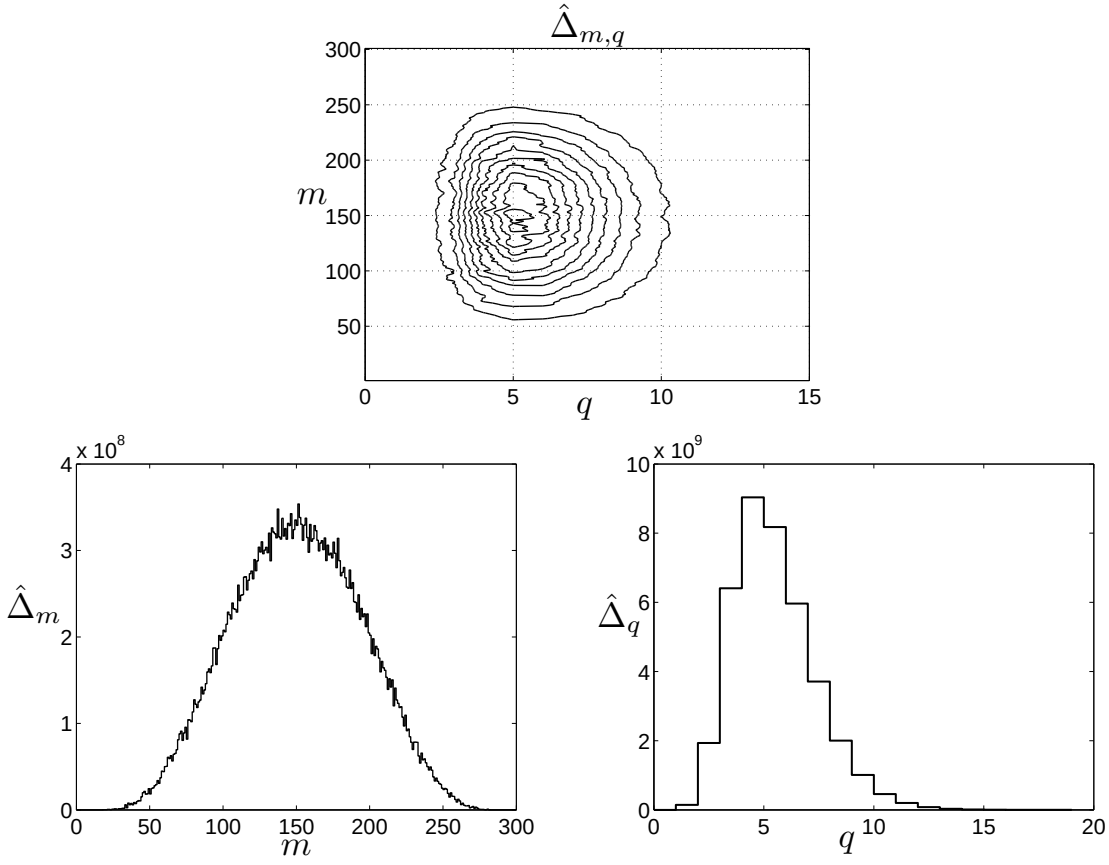

Рис. 6.1. Профили расслоения-связности семейства линейных классификаторов в $\mathbb{R}^p$: **Вверху**: изолинии поверхности профиля $\tilde{\Delta}_{m,q}$; **Внизу**: профиль расслоения $\tilde{\Delta}_m$ и профиль связности $\tilde{\Delta}_q^+$; $d = 5,\ |\mathbb{X}| = 300,\ |\hat{A}| = 2 \cdot 10^5$.

Процедура Π_A:

1. Случайно выбирается одно из $\binom{L}{d}$ подмножеств $S \subset \mathbb{X}, |S| = d$;
2. Проводится гиперплоскость $\tilde{w}$ через точки S, и случайно выбирается одна из ее 2 возможных ориентаций; это дает нам классификацию точек $\mathbb{X} \setminus S$;
3. Случайно выбирается одна из 2^d возможных классификаций точек S;
4. Вектор $\tilde{w}$ изменяется на небольшую величину так, чтобы классификатор $a_{\tilde{w}}(x)$ давал выбранную классификацию точек S без изменения классификация точек $\mathbb{X} \setminus S$;
5. Рассчитывается вектор ошибок $\tilde{a} \in A$ классификатора $a_{\tilde{w}}(x)$.

На выходе процедуры Π_A имеем вектор ошибок $\tilde{a}$ и представляющую его гиперплоскость $\tilde{w}$.

Процедура Π_A выбирает алгоритмы неравномерно — различные алгоритмы в A имеют различную вероятность быть выбранными; определим эти вероятности.

Определение 6.2.1. *Пусть $\mathcal{F}$ есть семейство линейных классификаторов в $\mathbb{R}^d$ и $\mathbb{X}$ полная выборка. Тогда для алгоритма $a \in A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ назовём связующими подмножествами такие $S \subset \mathbb{X}, |S| = d$, что проведенная через S и должным образом ориентированная гиперплоскость дает ту же классификацию точек $\mathbb{X} \setminus S$, что и a. Обозначим $\rho_d(a)$ число связующих подмножеств для алгоритма a.*

Связующее подмножество обобщает понятие связующего объекта — очевидно, для каждого связующего подмножества S алгоритма a в A есть $2^d - 1$ алгоритмов, отличающиеся от a только на объектах S всеми возможными способами.

Тогда несложно видеть, что

Утверждение 6.2.2. *Вероятность выбора $a \in A$ процедурой Π_A есть*

$$p(a) \stackrel{\text{def}}{=} \frac{\rho_d(a)}{\binom{L}{d}} \cdot \frac{1}{2} \cdot \frac{1}{2^d}.$$

Для вычисления $\rho_d(\tilde{a})$ и $\rho_+(\tilde{a})$ заданного алгоритма $\tilde{a} \in \tilde{A}$ имеем следующую лемму.

Лемма 6.2.3. *Пусть $\mathcal{F}$ есть семейство* (5.1) *линейных классификаторов в $\mathbb{R}^d$, пусть* I *есть индикатор ошибки* (5.2)*, пусть полная выборка $\mathbb{X} \subset \mathbb{R}^d$ находится в общем положении и пусть $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$ есть множество алгоритмов. Для алгоритма $\tilde{a} \in A$ пусть $a_{\tilde{w}}(x) \in \mathcal{F}$ есть произвольный представляющий его классификатор. Тогда связующие объекты алгоритма $\tilde{a}$ соответствуют вершинам выпуклой оболочки:*

$$\operatorname{conv}\left\{ \frac{x}{\langle \tilde{w}, x \rangle} : \; x \in \mathbb{X} \right\},$$

а его связующие множества соответствуют d-мерным граням оболочки.

Доказательство. Обозначим $\tilde{y}_i = \operatorname{sign}\langle \tilde{w}, x_i \rangle, \; x_i \in \mathbb{X}$ ответы $\tilde{a}$ на $\mathbb{X}$. Отметим, что $\tilde{y}_i \langle \tilde{w}, x_i \rangle > 0$ и $\langle \tilde{w}, x_i \rangle \neq 0$.

Рассмотрим конфигурацию L гиперплоскостей $\{h(x) \colon x \in \mathbb{X}\}$ в пространстве $\mathbb{W} \equiv \mathbb{R}^{p+1}$ параметров семейства $\mathcal{F}$: $h(x) = \{w \in \mathbb{W} \colon \langle w, x \rangle = 0\}, x \in \mathbb{X}$. Ячейка конфигурации $\tilde{c} \subset \mathbb{W}$, соответствующая $\tilde{a}$, есть $\tilde{c} = \{w \colon \tilde{y}_i \langle w, x_i \rangle \geqslant 0, \; \forall x_i \in \mathbb{X}\}$. Очевидно, что грани $\tilde{c}$ соответствуют связующим объектам $\tilde{a}$, вершины $\tilde{c}$ соответствуют связующим подмножествам $\tilde{a}$; $\tilde{w}$ есть внутренняя точка $\tilde{c}$.

Заметим, что трансляция координат не изменит структуру $\tilde{c}$ и сдвинем начало координат $w \mapsto w + \tilde{w}$ в точку $\tilde{w}$ так, чтобы начало координат стало внутренней точкой $\tilde{c}$. В новых координатах $\tilde{c} = \{w \colon \tilde{y}_i \langle w, x_i \rangle + \tilde{y}_i \langle \tilde{w}, x_i \rangle \geqslant 0, \; \forall x_i \in \mathbb{X}\}$, деля каждое неравенство в последнем выражении на $\tilde{y}_i \langle \tilde{w}, x_i \rangle \; (> 0)$, имеем $\tilde{c} = \left\{ w \colon \langle w, \frac{x_i}{\langle \tilde{w}, x_i \rangle} \rangle \leqslant 1, \forall x_i \in \mathbb{X} \right\}$. Несложно показать, что двойственный многогранник для $\tilde{c}$, записанного в такой форме, есть $\tilde{c}^* = \operatorname{conv}\left\{ \frac{x_i}{\langle \tilde{w}, x_i \rangle} \colon x_i \in \mathbb{X} \right\}$ (см. к примеру, [31]);

по определению двойственного многогранника, существует биективное соответствие между вершинами $\tilde{c}^*$ и гранями $\tilde{c}$, а также между гранями $\tilde{c}^*$ и вершинами $\tilde{c}$. ■

Для подсчета граней выпуклой оболочки существуют достаточно быстрые алгоритмы, к примеру, Qhull [15].

Обозначим $p_{un} = \frac{1}{|A|}$ вероятность выбора a в случае равномерного выбора. Тогда, очевидно, для получения несмещенных оценок профилей мы должны учитывать каждый выбранный алгоритм $\tilde{a} \in \tilde{A}$ с весом $\frac{p_{un}}{p(\tilde{a})}$:

Утверждение 6.2.4. *Если подмножество алгоритмов $\tilde{A} \subset A$ выбрано процедурой Π_A, то несмещенные оценки профилей расслоения-связности есть*

$$\tilde{\Delta}_{m,q} = \frac{|A|}{|\tilde{A}|} \cdot \sum_{\tilde{a}\in\tilde{A}} \frac{p_{un}}{p(\tilde{a})} \left[n_{\tilde{a}} = m \right] \left[\rho_+(\tilde{a}) \right],$$
$$\tilde{\Delta}_m = \frac{|A|}{|\tilde{A}|} \cdot \sum_{\tilde{a}\in\tilde{A}} \frac{p_{un}}{p(\tilde{a})} \left[n_{\tilde{a}} = m \right],$$
$$\tilde{\Delta}_s = \frac{|A|}{|\tilde{A}|} \cdot \sum_{\tilde{a}\in\tilde{A}} \frac{p_{un}}{p(\tilde{a})} \left[\hat{n}_{\tilde{a}} = s \right].$$

6.3 Вычисление оценок

Вычисление Shell-оценок (теоремы 3.3.1, 3.3.3, 3.4.3) и оценок, учитывающих сходство алгоритмов (теоремы 4.4.3, 4.5.7, 4.6.3), производится по следующей схеме:

1. Из смеси двух нормальных распределений в $\mathbb{R}^5$ (с байесовским уровнем ошибки 15%) генерируется случайная двухклассовая полная выборка $\mathbb{X} \subset \mathbb{R}^5$ размера $L = 300$.
2. Делается сэмплинг $|\tilde{A}| = 100.000$ алгоритмов из множества $A = \mathrm{I}(\mathcal{F}, \mathbb{X})$.
3. Рассчитываются оценки профилей $\Delta_{m,q}, \Delta_m, \hat{\Delta}_s$ (последний – на случайной обучающей выборке $X \subset \mathbb{X}$ длины $\ell = 100$).
4. По оценкам профилей вычисляются оценки теорем 3.3.1, 3.3.3, 3.4.3, 4.4.3, 4.5.7, 4.6.3 и VC-оценка Теоремы 1.5.1.
5. Делается сэмплинг $|\mathfrak{X}| = 100.000$ обучающих выборок из $[\mathbb{X}]^\ell$, на них производится обучение методом μ (SVM) и получается оценка распределения $P(m, s)$, определенного в (1.17).
6. По оценке распределения $P(m, s)$ вычисляется оценка Леммы 1.7.1.

Пусть $\bar{\nu}\left(\hat{\nu}, \eta(\alpha)\right)$ — любая из перечисленных выше доверительных оценок для истинной частоты ошибок, вычисленная при доверительном уровне α, где $\eta(\alpha)$ —

соответствующий порог, либо набор порогов переобучения. Определим разность

$$\varepsilon(\hat{\nu}, \alpha) \stackrel{\text{def}}{=} \bar{\nu}\left(\hat{\nu}, \eta(\alpha)\right) - \hat{\nu},$$

имеющую смысл верхней оценки для уклонения истинной частоты от наблюдаемой частоты. Например, на Рис. 1.3 эта разность соответствует горизонтальному расстоянию от диагонали до границы, соответствующей области переобучения.

6.3.1 Shell-оценки

На Рис. 6.2 приведены кривые $\varepsilon(\hat{\nu}, \alpha)$ для Shell-оценок, вычисленных при фиксированной вероятности переобучения $\alpha = 5\%$.

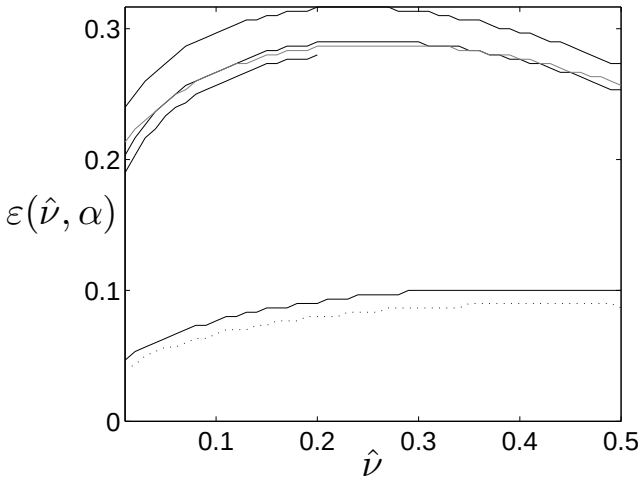

Рис. 6.2. Точность различных оценок $\bar{\nu}$ истинной частоты ошибок в зависимости от наблюдаемой частоты ошибок $\hat{\nu}$ при фиксированном доверительном уровне $\alpha = 0.05$.

- Самая нижняя кривая (пунктир) соответствует оценке $\bar{n}\left(\hat{n}, \alpha\right)$ для одного алгоритма.

- Кривая над ней соответствует $\bar{n}\left(\hat{n}, \eta_{exact}(\alpha)\right)$ — точной доверительной оценке для метода SVM.

- Серая кривая соответствует VC-оценке $\bar{n}\left(\hat{n}, \frac{\alpha}{|A|}\right)$.

- Чёрная кривая соответствует shell-оценке $\bar{n}\left(\hat{n}, \eta(\hat{n}, \alpha)\right)$ Теоремы 3.3.3. Как и следовало ожидать, она несущественно отличается от VC-оценки. Она чуть точнее VC-оценки в области малых и больши́х частот и менее точна в средней области.

- Короткая кривая под ней соответствует оценке $\bar{n}\,(\hat{n}, \eta(\hat{n}, \alpha))$ Следствия 3. Она показывает, в некотором смысле, максимальное улучшение, которого можно добиться от shell-оценки Теоремы 3.3.3 для данного профиля Δ_m за счет учета информации о распределении $P(\hat{\nu})$.

- Самая верхняя кривая соответствует shell-оценке $\bar{n}\,(\hat{n}, \hat{\eta}(\hat{n}, \alpha))$ Теоремы 3.4.3, вычисляемой по профилю наблюдаемых частот; она по определению хуже shell-оценки, вычисляемой по Δ_m, и, в данном случае, равномерно хуже VC-оценки.

Отметим, что все оценки более точны для малых значений $\hat{\nu}$, что является следствием использования квантильного условия переобучения. Доверительная VC-оценка, сделанная на основе условия $\nu - \hat{\nu} \geqslant \varepsilon$, будет на этом графике горизонтальной прямой, проходящей по верхней границе серой кривой.

Как можно видеть, shell-оценки не дают практически никакого улучшения в сравнении с VC-оценкой. Это является следствием того, что обе shell-оценки (Теоремы 3.3.1, 3.3.3 и, соответственно, [49], [47]) наследуют два основных фактора завышенности VC-оценки — оценку равномерного уклонения частот и неравенство Буля. Фактически, обе shell-оценки являются не более чем различными способами перегруппировки слагаемых гипергеометрической вероятности $h_m(s)$ в VC-оценке, поэтому не могут быть существенно точнее последней.

6.3.2 Оценки с использованием связности

На Рис. 6.3 приведены кривые $\varepsilon(\hat{\nu}, \alpha)$ для оценок, учитывающих связность семейства и вычисленных при фиксированной вероятности переобучения $\alpha = 5\%$.

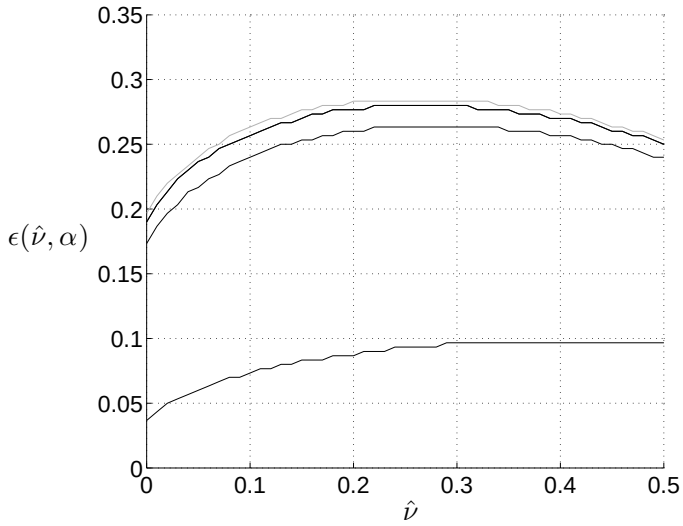

Рис. 6.3. Зависимости $\varepsilon(\hat{\nu}, \alpha)$ от $\hat{\nu}$.

- Самая нижняя кривая соответствует точной доверительной оценке для метода SVM.

- Самая верхняя (серая) кривая соответствует VC-оценке $\bar{n}\,(\hat{n}, \frac{\alpha}{|A|})$.

- Кривая непосредственно под VC-оценкой соответствует практически совпадающим оценкам $\bar{n}\,(\hat{n}, \eta_{tree}(\alpha))$ и $\bar{n}\,(\hat{n}, \eta_{chains}(\alpha))$ теорем 4.4.3 и 4.6.3.

- Кривая под ней соответствует оценке $\bar{n}\,(\hat{n}, \eta_{conn}(\alpha))$ теоремы 4.5.7.

Как можно видеть, учет одной монотонной цепи, исходящей из каждого алгоритма, не дает практически никакого улучшения оценки по сравнению с учетом одной связи алгоритма. В то же время, учет полустепени связности алгоритма дает значимое улучшение оценки, хотя она по-прежнему остается завышенной в сравнении с точной оценкой по методу Монте-Карло. Все три оценки дают большее улучшение, чем shell-оценки предыдущего параграфа.

Интересно также отметить, что оценка, учитывающая полустепень связности, уменьшается экспоненциально в сравнении с VC-оценкой с ростом размерности пространства параметров. Отношения оценок в зависимости от размерности приведены на Рисунке 6.4.

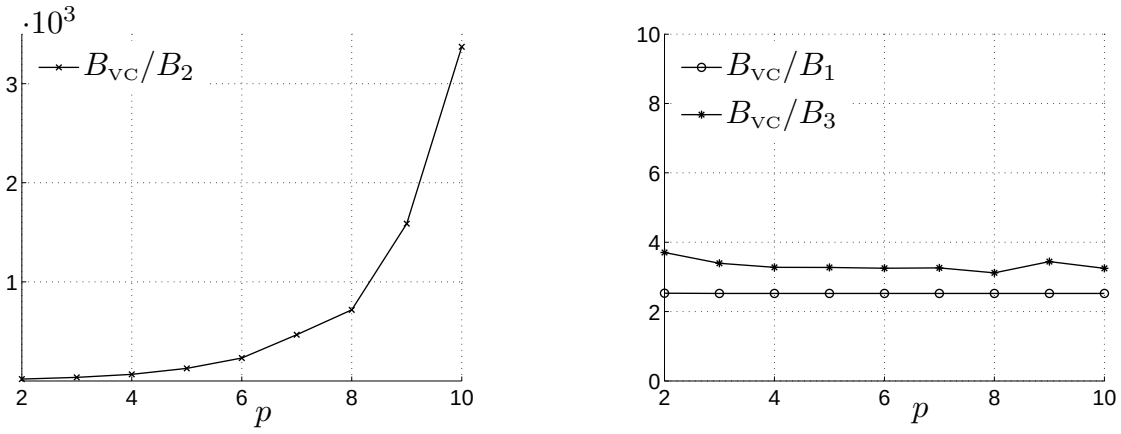

Рис. 6.4. Отношение оценки неравенства Буля к оценкам для семейства $\mathcal{F}$ линейных классификаторов и случайной полной выборки $\mathbb{X} \subset \mathbb{R}^{d-1}$, при $|\mathbb{X}| = 300$, $\ell = 100$, $|\hat{A}| = 10.000$:
$B_{\rm VC} = \eta\,|A|$ — VC-оценка (Теорема 1.5.1);
B_1 — оценка, использующая связность A (Теорема 4.4.3);
B_2 — оценка, использующая профиль расслоения-связности (Теорема 4.5.7);
B_3 — оценка, использующая цепи алгоритмов (Теорема 4.6.3);

Уже при небольших размерностях d оценка B_2 вероятности переобучения на несколько порядков лучше, чем VC-оценка, и их отношение растет с размерностью экспоненциально как $\approx 2^d$. Оценки B_1, B_3 лишь в несколько раз лучше VC-оценки независимо от d.

Заключение

Результаты, выносимые на защиту:

1. Получены комбинаторные аналоги shell-оценок Лэнгфорда-МакАллистера, показано, что shell-оценки являются аналогом стандартных оценок Вапника-Червоненкиса и «бритвы Оккама» Блумера и имеют ту же степень завышенности.

2. Предложен новый способ учета сходства алгоритмов в оценках вероятности переобучения, основанный на Бонферрони-оценках вероятности равномерного отклонения частот и методе производящих функций.

3. Получены оценки вероятности переобучения для случаев связного семейства, семейства с известным профилем расслоения-связности, семейства, состоящего из множества монотонных максимальных цепей алгоритмов.

4. Для семейства линейных классификаторов получены оценки среднего и дисперсии профиля связности.

Список обозначений

$\langle w, x\rangle$ — скалярное произведение векторов $w, x \in \mathbb{R}^p$;

$\vee$ — дизъюнкция предикатов;

$\wedge$ — конъюнкция предикатов;

$\stackrel{\text{def}}{=}$ — определение величины слева от равенства;

$[\mathbb{X}]^\ell$ — множество всех ℓ-элементных подмножеств множества $\mathbb{X}$;

$\operatorname{card}\{X\}$ — мощность множества X;

$[\mathrm{U}] = \begin{cases} 0, & \text{U=ложь;} \\ 1, & \text{U=истина;} \end{cases}$ — индикаторная (характеристическая) функция предиката U;

$\lceil x \rceil$ — функция «потолок» — минимальное целое, не меньшее x;

$\lfloor x \rfloor$ — функция «пол» — максимальное целое, не большее x;

$\arg\min\limits_{s \in S} f(s)$ — произвольный элемент из множества точек минимума;

A — множество различных бинарных векторов ошибок функций из $\mathcal{F}$ на полной выборке $\mathbb{X}$, в работе для краткости также называется множеством алгоритмов;

A_m — m-й слой множества алгоритмов A;

$A_\pm(a)$ — «единичная окрестность» a — множество алгоритмов в A отличающихся от $a \in A$ на одном объекте;

A_1^2 — множество ребер графа G_A^1;

$a = \big(I(a, x_i)\big)_{i=1}^L$ — вектор ошибок алгоритма $a \in A$;

$\hat{a} \equiv \mu X$ — алгоритм $\mu(X)$, получаемый в результате обучения по выборке X;

$C_n^k = \frac{n!}{k!\,(n-k)!}$ — биномиальные коэффициенты;

cl — топологическое замыкание множества;

$\operatorname{conv}\{\mathbb{X}\}$ — выпуклая оболочка множества $\mathbb{X}$ когда $\mathbb{X} \subset \mathbb{R}^p$;

$\mathcal{C}$ — множество ячеек конфигурации гиперплоскостей $\mathcal{H}$;

Δ_m — профиль расслоения множества алгоритмов A — число алгоритмов с m ошибками на полной выборке;

$\Delta_{m,q}$ — профиль расслоения и связности множества алгоритмов;

Δ_q^+ — профиль связности множества алгоритмов A — число алгоритмов имеющих ровно q «верхних соседей» в графе 1-сходства A;

$\tilde{\Delta}_{m,q}, \tilde{\Delta}_m, \tilde{\Delta}_q^+$ — оценки профилей $\Delta_{m,q}, \Delta_m, \Delta_q^+$ по методу Монте-Карло на основе выборки $\tilde{A} \subset A$ из множества алгоритмов;

$\hat{\Delta}_s$ — наблюдаемый профиль расслоения (или профиль наблюдаемых частот) множества A — число алгоритмов с s ошибками на заданной обучающей выборке;

ε — порог переобучения в критерии переобучения $\nu_a - \hat{\nu}_a \geqslant \varepsilon$;

η — порог переобучения в критерии переобучения $H_{n_a}(\hat{n}_a) \leqslant \eta$;

$\mathbf{E}$ — матожидание как среднее по множеству обучающих выборок $\frac{1}{C_L^\ell} \sum_{X \in [\mathbb{X}]^\ell}$;

$\mathcal{F}$ — множество алгоритмов, параметрическое семейство функций из $\mathbb{X}$ в Y;

$\mathcal{F}$ — множество граней конфигурации гиперплоскостей $\mathcal{H}$;

G_A^1 — граф 1-сходства, в котором вершины — алгоритмы из множества A, ребрами соединены алгоритмы отличные на одном объекте;

$h_m(s) = \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ — гипергеометрическое распределение (ГГР);

$H_m(s) = \sum\limits_{t=0}^{\lfloor s \rfloor} h_m(t)$ — функция распределения ГГР;

$\mathcal{H}$ — конфигурация L однородных гиперплоскостей в $\mathbb{R}^{p+1}$ индуцированная множеством $\mathbb{X} \subset \mathbb{R}^p$;

$I(a,x) = \big[$алгоритм a ошибается на объекте $x\big]$; — индикатор ошибки

$\mathrm{I}(\mathcal{F}, \mathbb{X})$ — множество A бинарных векторов ошибок на полной выборке $\mathbb{X}$, индуцированное множеством функций $\mathcal{F}$ и функцией индикатора ошибки I;

ℓ — длина наблюдаемой (обучающей) выборки X;

L — длина генеральной совокупности (или *полной выборки*) $\mathbb{X}$;

$\mu\colon 2^{\mathbb{X}} \to A$ — метод обучения;

m — обычно число ошибок на генеральной выборке, $m = n(a, \mathbb{X})$;

$\mathbf{M}(a_1, \ldots, a_k)$ — бинарная матрица со строками — векторами ошибок алгоритмов $a_1, \ldots, a_k$;

$n_a = n(a, X)$ — число ошибок алгоритма $a \in A$ на выборке $X \subseteq \mathbb{X}$;

$\nu_a = \nu(a, X) = \frac{1}{|X|} n(a, X)$ — частота ошибок алгоритма $a \in A$ на выборке $U \subseteq \mathbb{X}$;

$\hat{n}_a$ — число ошибок алгоритма a на заданной обучающей выборке $\bar{X}$;

$\hat{\nu}_a$ — частота ошибок алгоритма a на заданной обучающей выборке $\bar{X}$;

$\bar{n}\left(\hat{n}_a, \eta\right)$ — доверительная верхняя граница для числа ошибок алгоритма a на полной выборке $\mathbb{X}$ при уровне надежности η и наблюдаемом числе ошибок $\hat{n}_a$ на обучающей выборке X;

$\mathbb{X}_{\pm}(a)$ — множество «связующих объектов» объектов для алгоритма a в $\mathbb{X}$;

$\mathbf{P} = \frac{1}{C_L^\ell} \sum\limits_{X \in [\mathbb{X}]^\ell}$ — вероятность как доля разбиений генеральной выборки;

$P(a)$ — вероятность получить алгоритм a в результате обучения;

Q_A — функционал равномерного отклонения частот;

$\rho_{\pm}(a)$ — верхняя/нижняя полустепень связности алгоритма a;

$\rho(a, a')$ — расстояние Хэмминга между векторами ошибок алгоритмов a, a';

$\mathbb{R}$ — множество действительных чисел;

s — обычно число ошибок на обучающей выборке, $s = n(a, X)$;

$\operatorname{sign}(x) = \begin{cases} -1, & x<0 \\ +1, & x>0 \end{cases}$ — функция «знак»;

s_m — граница переобучения в записи критерия переобучения $\hat{n}_a \leqslant s_{n_a}$; для квантильного критерия переобучения $s_{n_a}(\eta) = \max\{s\colon H_{n_a}(s) \leqslant \eta\}$;

$\hat{\sigma}$ — бинарный случайный вектор с компонентами соответствующими объектам в $\mathbb{X}$ в котором $\hat{\sigma}_i = 1$, если $x_i \in X$ и $\hat{\sigma}_i = 0$ иначе;

U_a — предикат «алгоритм a является переобученным для обучающей выборки X»;

Var — дисперсия случайной величины;

$\mathbb{W}$ — пространство $\mathbb{R}^{p+1}$ параметров семейства $\mathcal{F}$ линейных классификаторов в $\mathbb{R}^p$;

$\mathbb{X} = \{x_1, \ldots, x_L\}$ — полная выборка;

X — наблюдаемая (обучающая) выборка;

$\bar{X}$ — скрытая (контрольная) выборка;

6.4 Список литературы

[1] *Бернштейн С.* Теория вероятностей. — Гаztехиздат, Москва, 1946.

[2] *Вапник В.* Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979.

[3] *Вапник В., Червоненкис А.* Теория распознавания образов. — Наука, 1974.

[4] *Воронцов К. В.* Комбинаторный подход к оценке качества обучаемых алгоритмов // *Мат. вопросы кибернетики.* — 2004. — Vol. 13. — Pp. 5–36.

[5] *Воронцов К. В.* Комбинаторная теория надёжности обучения по прецедентам. — Диссертация на соискание учёной степени д.ф.-м.н., М.: ВЦ РАН. — 2010. `http://www.MachineLearning.ru/wiki/images/b/b6/Voron10doct.pdf`.

[6] *Кочедыков Д. А.* Комбинаторные оценки обобщающей способности методов обучения по прецедентам с расслоением по наблюдаемой частоте ошибок // Труды 51-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук»: Часть VII. Управление и прикладная математика. — 2008.

[7] *Кочедыков Д. А.* Структуры сходства в семействах алгоритмов классификации и оценки обобщающей способности // Докл. 14-й Всеросс. конф. «Математические методы распознавания образов». — 2009. — Pp. 45–48.

[8] *Кочедыков Д. А., Воронцов К. В.* О поиске оптимальных сочетаний управляющих параметров в логических алгоритмах классификации // Тезисы докл. межд. конф. «Интеллектуализация обработки информации», Симферополь. — 2006. — Pp. 117–118.

[9] *Кочедыков Д. А., Воронцов К. В.* К определению понятия информативности логических закономерностей в задачах классификации // Труды 50-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук»: Часть VII. Управление и прикладная математика. — 2007. — Pp. 279–281.

[10] *Кочедыков Д. А., Воронцов К. В., Ивахненко А. А.* Система кредитного скоринга на основе логических алгоритмов классификации // Докл. 12-й Всеросс. конф. «Математические методы распознавания образов». — 2005. — Pp. 349–353.

[11] *Кочедыков Д. А., Воронцов К. В., Ивахненко А. А.* Применение логических алгоритмов классификации в задачах кредитного скоринга и управления риском кредитного портфеля банка // Докл. 13-й Всеросс. конф. «Математические методы распознавания образов». — 2007. — Pp. 484–488.

[12] *Agarwal P. K., Sharir M.* Arrangements and their applications // Handbook of Computational Geometry. — 1998. — Pp. 49–119.

[13] *Anthony M., Bartlett P.* Neural network learning: theoretical foundations. — Cambridge University Press, 1999.

[14] *Anthony M., Brightwell G., Shawe-Taylor J.* On specifying boolean functions by labelled examples // *Discrete Appl. Math.* — 1995. — Vol. 61, no. 1. — Pp. 1–25.

[15] *Barber C. B., Dobkin D., Huhdanpaa H.* The quickhull algorithm for convex hulls // *ACM Transactions on Mathematical Software.* — 1996. — Vol. 22, no. 4. — Pp. 469–483.

[16] *Björner A.* Oriented matroids. — Cambridge Univ. Press, 1999.

[17] *Boucheron S., Bousquet O., Lugosi G.* Concentration inequalities // Advanced lectures on machine learning. — Springer, 2004. — Pp. 208–240.

[18] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: some recent advances // *ESAIM Probability & Statistics.* — 2005. — Vol. 9. — Pp. 323–375.

[19] *Bousquet O., Boucheron S., Lugosi G.* Introduction to statistical learning theory // Advanced Lectures in Machine Learning. — Springer, 2004. — Pp. 169–207.

[20] *Bousquet O., Boucheron S., Lugosi G.* Introduction to statistical learning theory // Advanced Lectures in Machine Learning. — Springer, 2004. — Pp. 169–207.

[21] *Cristianini N., Shawe-Taylor J.* An introduction to support Vector Machines: and other kernel-based learning methods. — Cambridge University Press, 2000.

[22] *Devroye L., Györfi L., Lugosi G.* A Probabilistic Theory of Pattern Recognition. — Springer-Verlag, 1996.

[23] *Devroye L., Lugosi G.* Lower bounds in pattern recognition and learning // *Pattern Recognition.* — 1995. — Vol. 28, no. 7. — Pp. 1011–1018.

[24] *Diaz J., Petit J., Serna M.* A guide to concentration bounds // Handbook on randomized computing. — Vol. II. — Kluwer Press, 2001. — Pp. 457–507.

[25] *Dohmen K.* Improved Bonferroni Inequalities via Abstract Tubes. — Springer-Verlag, 2003.

[26] *Edelsbrunner H.* Algorithms in Combinatorial Geometry. — Springer-Verlag, 1987.

[27] *Edelsbrunner H., Seidel R., Sharir M.* On the zone theorem for hyperplane arrangements // New Results and New Trends in Computer Science. — Springer-Verlag, 1991. — Pp. 108–123.

[28] *Flajolet P., Sedgewick R.* Analytic combinatorics. — Cambridge University Press, 2009. `http://algo.inria.fr/flajolet/Publications/book.pdf`.

[29] *Galambos J., Simonelli I.* Bonferroni-type Inequalities with Applications. — Springer-Verlag, 1996.

[30] *Grunbaüm B.* Convex Polytopes. — Springer, 2003.

[31] *Grünbaum B., Klee V.* Convex Polytopes. — Springer-Verlag, 2003.

[32] *Halperin D.* Arrangements // Handbook of discrete and computational geometry / Ed. by J. O. Jacob E. Goodman. — Chapman and Hall, 2004.

[33] *Haussler D.* Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension // *J. Comb. Theory Ser. A.* — 1995. — Vol. 69, no. 2. — Pp. 217–232.

[34] *Haussler D., Littlestone N., Warmuth M.* Predicting (0, 1)-functions on randomly drawn points // *Foundations of Computer Science, Annual IEEE Symposium on.* — 1988. — Vol. 0. — Pp. 100–109.

[35] *Herbrich R.* Learning Kernel Classifiers: Theory and Algorithms. — MIT Press, 2001.

[36] *Hoeffding W.* Probability inequalities for sums of bounded random variables // *Journal of the American Statistical Association.* — 1963. — Vol. 58, no. 301. — Pp. 13–30.

[37] *Hunter D.* An upper bound for the probability of a union // *J. Appl.Probab.* — 1976. — Vol. 13. — Pp. 597–603.

[38] *Hush D., Scovel C.* Concentration of hypergeometric distribution // *Statistics & Probability Letters.* — 2005. — Vol. 75, no. 2. — Pp. 127–132.

[39] *Johnson N. L., Kotz S., Kemp A. W.* Univariate Discrete Distributions. — Wiley-Interscience Publication, 1992.

[40] *Kearns M. J., Schapire R. E.* Efficient distribution-free learning of probabilistic concepts // Computational Learning Theory and Natural Learning Systems, Volume I: Constraints and Prospect, edited by Stephen Jose Hanson, George A. Drastal, and Ronald L. Rivest, Bradford/MIT Press. — 1994. — Vol. 1. `citeseer.ist.psu.edu/article/kearns93efficient.html`.

[41] *Kochedykov D. A.* Combinatorial shell bounds for generalization ability // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20. — Pp. 459–473.

[42] *Kochedykov D. A.* A combinatorial approach to hypothesis similarity in generalization bounds // *Pattern Recognition and Image Analysis.* — 2011. — Vol. 21.

[43] *Kolmogorov A. N., Tikhomirov V. M.* ε-entropy and ε-capacity of sets in function spaces // *Translations of the American Math. Soc.* — 1961.

[44] *Koltchinskii V.* Rademacher penalties and structural risk minimization // *IEEE Transactions on Information Theory.* — 2001. — Vol. 47, no. 5. — Pp. 1902–1914. `citeseer.ist.psu.edu/391084.html`.

[45] *Koltchinskii V., Panchenko D.* Rademacher processes and bounding the risk of function learning // High Dimensional Probability, II / Ed. by D. E. Gine, J.Wellner. — Birkhauser, 1999. — Pp. 443–457. `citeseer.ist.psu.edu/article/koltchinskii99rademacher.html`.

[46] *Koltchinskii V., Panchenko D.* Empirical margin distributions and bounding the generalization error of combined classifiers // *The Annals of Statistics.* — 2002. — Vol. 30, no. 1. — Pp. 1–50.

[47] *Langford J.* Quantitatively tight sample complexity bounds: Ph.D. thesis / Carnegie Mellon Univ. — 2002.

[48] *Langford J., McAllester D.* Computable shell decomposition bounds // Proc. 13th Annu. Conference on Comput. Learning Theory. — Morgan Kaufmann, San Francisco, 2000. — Pp. 25–34. citeseer.ist.psu.edu/langford00computable.html.

[49] *Langford J., McAllester D.* Computable shell decomposition bounds // *J. Mach. Learn. Res.* — 2004. — Vol. 5. — Pp. 529–547.

[50] *Langford J., Shawe-Taylor J.* PAC-Bayes and margins // Advances in Neural Information Processing Systems 15. — MIT Press, 2002. — Pp. 439–446.

[51] *Linhart J.* Arrangements of oriented hyperplanes // *Discrete and Computational Geometry.* — 1993. — Vol. 10, no. 1. — Pp. 435–446.

[52] *McDiarmid C.* On the method of bounded differences // *In Surveys in Combinatorics, London Math. Soc. Lecture Notes Series.* — 1989. — Vol. 141. — Pp. 148–188.

[53] *Mendelson S.* A few notes on statistical learning theory // Advanced lectures on machine learning. — Springer-Verlag New York, Inc., 2003. — Pp. 1–40.

[54] *Naiman D. Q., Wynn H. P.* Abstract tubes, improved inclusion-exclusion identities and inequalities and importance sampling // *The Annals of Statistics.* — 1997. — Vol. 25, no. 5. — Pp. 1954–1983.

[55] Occam's razor / A. Blumer, A. Ehrenfeucht, D. Haussler, M. K. Warmuth // *Inf. Process. Lett.* — 1987. — Vol. 24. — Pp. 377–380.

[56] *Peter Orlik H. T.* Arrangements of Hyperplanes. — Springer, 2010.

[57] *Pollard D.* Convergence of stochastic processes. — Springer-Verlag, 1984.

[58] Rigorous learning curve bounds from statistical mechanics / D. Haussler, M. Kearns, H. S. Seung, N. Tishby // *Machine Learning.* — 1996. — no. 25. — Pp. 195–236.

[59] Scale-sensitive dimensions, uniform convergence, and learnability / N. Alon, S. Ben-David, N. Cesa-Bianchi, D. Haussler // *J. ACM.* — 1997. — Vol. 44. — Pp. 615–631.

[60] *Sheffer T., Joachims T.* Expected model analisys for model selection // International conference on model selection. — 1999.

[61] *Sill J.* Monotonicity and connectedness in learning systems: Ph.D. thesis / CalTech. Univ. — 1998.

[62] *Stanley R. P.* An introduction to hyperplane arrangements // Geometric Combinatorics / Ed. by V. R. E. Miller, B. Sturmfels. — 2007.

[63] *Valiant L. G.* A theory of the learnable // Proceedings of the sixteenth annual ACM symposium on Theory of computing. — ACM, 1984. — Pp. 436–445.

[64] *Vapnik V. N.* Statistical Learning Theory. — Wiley, 1998.

[65] *Vapnik V. N., Chervonenkis A. Y.* On the uniform convergence of relative frequencies of events to their probabilities // *Theory Probab. Appl.* — 1971. — Vol. 16. — Pp. 264–280.

[66] *Wilf H.* Generatingfunctionology. Ak Peters Series. — A K Peters, 2006. www.math.upenn.edu/~wilf/gfologyLinked2.pdf.